Uncovering the Transcription Factor Network Underlying Mammalian Sex Determination

by

Anirudh Natarajan

Program in Computational Biology and Bioinformatics
Duke University

Date:_____

Approved:

_____
Blanche Capel, co-supervisor

_____
Uwe Ohler, co-supervisor

_____
Gregory E. Crawford

_____
L. Ryan Baugh

_____
David R. McClay

_____
Terrence S. Furey

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in the Program in Computational Biology and Bioinformatics
in the Graduate School of Duke University
2014

ABSTRACT

Uncovering the Transcription Factor Network Underlying Mammalian Sex Determination

by

Anirudh Natarajan

Program in Computational Biology and Bioinformatics
Duke University

Date:_____

Approved:

_____
Blanche Capel, co-supervisor

_____
Uwe Ohler, co-supervisor

_____
Gregory E. Crawford

_____
L. Ryan Baugh

_____
David R. McClay

_____
Terrence S. Furey

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in the Program in Computational Biology and Bioinformatics
in the Graduate School of Duke University
2014

Abstract

Understanding transcriptional regulation in development and disease is one of the central questions in modern biology. The current working model is that Transcription Factors (TFs) combinatorially bind to specific regions of the genome and drive the expression of groups of genes in a cell-type specific fashion. In organisms with large genomes, particularly mammals, TFs bind to enhancer regions that are often several kilobases away from the genes they regulate, which makes identifying the regulators of gene expression difficult. In order to overcome these obstacles and uncover transcriptional regulatory networks, we used an approach combining expression profiling and genome-wide identification of enhancers followed by motif analysis. Further, we applied these approaches to uncover the TFs important in mammalian sex determination.

Using expression data from a panel of 19 human cell lines we identified genes showing patterns of cell-type specific up-regulation, down-regulation and constitutive expression. We then utilized matched DNase-seq data to assign DNase Hypersensitivity Sites (DHSs) to each gene based on proximity. These DHSs were scanned for matches to motifs and compiled to generate scores reflecting the presence of TF binding sites (TFBSs) in each gene's putative regulatory regions. We used a sparse logistic regression classifier to classify differentially regulated groups of genes. Comparing our approach to

proximal promoter regions, we discovered that using sequence features in regions of open chromatin provided significant performance improvement. Crucially, we discovered both known and novel regulators of gene expression in different cell types. For some of these TFs, we found cell-type specific footprints indicating direct binding to their cognate motifs.

The mammalian gonad is an excellent system to study cell fate determination processes and the dynamic regulation orchestrated by TFs in development. At embryonic day (E) 10.5, the bipotential gonad initiates either testis development in XY embryos, or ovarian development in XX embryos. Genetic studies over the last 3 decades have revealed about 30 genes important in this process, but there are still significant gaps in our understanding. Specifically, we do not know the network of TFs and their specific combinations that cause the rapid changes in gene expression observed during gonadal fate commitment. Further, more than half the cases of human sex reversal are as yet unexplained.

To apply the methods we developed to identify regulators of gene expression to the gonad, we took two approaches. First, we carried out a careful dissection of the transcriptional dynamics during gonad differentiation in the critical window between E11.0 and E12.0. We profiled the transcriptome at 6 equally spaced time points and developed a Hidden Markov Model to reveal the cascades of transcription that drive the differentiation of the gonad. Further, we discovered that while the ovary maintains its

transcriptional state at this early stage, concurrent up- and down-regulation of hundreds of genes are orchestrated by the testis pathway. Further, we compared two different strains of mice with differential susceptibility to XY male-to-female sex reversal. This analysis revealed that in the C57BL/6J strain, the male pathway is delayed by ~5 hours, likely explaining the increased susceptibility to sex reversal in this strain. Finally, we validated the function of *Lmo4*, a transcriptional co-factor up-regulated in XY gonads at E11.6 in both strains. RNAi mediated knockdown of *Lmo4* in primary gonadal cells led to the down-regulation of male pathway genes including key regulators such as *Sox9* and *Fgf9*.

To find the enhancers in the XY gonad, we conducted DNase-seq in E13.5 XY supporting cells. In addition, we conducted ChIP-seq for H3K27ac, a mark correlated with active enhancer activity. Further, we conducted motif analysis to reveal novel regulators of sex determination. Our work is an important step towards combining expression and chromatin profiling data to assemble transcriptional networks and is applicable to several systems.

Contents

List of Tables

List of Figures

Acknowledgements

I have been fortunate to have not one, but two, fantastic advisors. Without Uwe Ohler and Blanche Capel, my computational and developmental biology mentors, my time in graduate school would not have been half as enjoyable and fruitful. Crucially, they gave me the independence to follow the questions I was interested in. I am most thankful for their patience in letting me learn biological techniques, which I was entirely useless at when I started graduate school. Their commitment to training me broadly is something I will treasure and appreciate throughout my scientific career.

The Capel lab was my physical home and I will always appreciate the time, generosity and shenanigans in the lab. I remember rotating with the 'old guard' of Steve Munger, Jonah Cool, Danielle Maatouk, Lindsey Mork, Matt Cook, Tony DeFalco and Samantha Jameson and thinking that this would be an enriching and happy place to be. Special thanks to good friend and mentor, Steve Munger who not only tolerated my utter incompetence at the bench, but helped me develop as a scientist and was so helpful in making me feel at home in the lab. Thanks also to Danielle and Samantha who I was fortunate to work with closely. They really contributed to my intellectual development. To the new folk – Jason Garness, Ximena Bustamante, Allison Navis, Michael Czerwinski, Yi-Tzu Lin, James Robinson, Xiaoyu Xia and Rachel Williamson – you have all made the last four years thoroughly enjoyable. A special thanks to everyone's favorite, Iordan Batchvarov, who not only helped me immeasurably with managing my

colony but also helped me buy my first ever car! While I did not sit in the Ohler lab, I gained a lot from being part of the lab and lab meetings. Special thanks to Gurkan who I worked with closely and was a very good sounding board for ideas. I was fortunate enough to meet the fantastic new group Uwe is putting together in Berlin and I am excited to see what the future brings from them. Big thanks to Neel Mukherjee who helped me settle in during my Berlin visit.

In addition to my advisors, I also have a wonderful committee– Greg Crawford, Ryan Baugh, Dave McClay, Terry Furey – who have taken the time and effort to guide me in my research. I was lucky to work with Greg on the DNase-seq work and appreciate his input in my scientific development. I also have to thank Ryan for his fantastic seminar course – Genes and Development - in my second year of graduate school.

The CBB community has made my time at Duke very enjoyable and I have made some dear friends over the years. Thanks to Liz Labriola for keeping me in line and helping me meet all the very important program deadlines. Among other events, I will also miss the nights filled with liar's dice and insult hurling at the CBB retreat.

My family has been extremely important in shaping the person that I have become. Thanks to my parents for being in awe of virtually all things in nature and passing that on to me. Thanks to my dad for engaging me in conversations about science, subscribing to Scientific American and having books about relativity and the

cosmic background radiation in the house. They made for a meaningful and wonderful upbringing.

To Lydia, my wife, I am so touched and thankful for the support, love and understanding over the last five years of graduate school. I cannot thank her enough for bearing 3 years of a long distance relationship, so I could go chase my dream. Without her, I would not have come to Duke, and without her, I definitely would not be where I am today.

1. Introduction

1.1 Transcriptional Gene Regulation

Since the advent of molecular biology in the mid 20th century, one of the central questions and major thrusts of research has been to understand the mechanisms of how cells achieve the widely varying and dynamic profiles of mRNA expression that are observed in a plethora of situations including response to the environment, embryonic development, and disease states. Our understanding of this fundamental process has advanced by leaps and bounds in the last few decades. Before delving into our modern day conception of the players involved in transcriptional regulation, it is illuminating to briefly consider a historical view of how this understanding has evolved in the recent past.

Groundbreaking work in understanding that non-coding DNA was important in directing gene expression came from work done in bacteria by Francois Jacob, Jacques Monod and Arthur Pardee (Darnell 2011). Specifically, by using bacterial genetics, they found operator sites, now known as Transcription Factor Binding Sites (TFBSs) that were bound by proteins, now known as Transcription Factors (TFs) to repress gene expression. These experiments opened the door to a major avenue into how organisms use a fixed genome to produce varied behaviors. Following this work, the recognition of the crucial role of messenger RNA (mRNA), and the identification of RNA polymerases, a key step came when the proteins responsible for repression were purified (Gilbert and Muller-Hill 1966; Ptashne 1967). Crucially, it was discovered that in these bacterial

systems, both activators and repressors act on gene regulation by interfering with the binding or interaction between RNA polymerase and other proteins that were part of the transcriptional machinery.

The situation in eukaryotic cells turned out to be far more complex. Using the technique of *in vitro* or cell-free transcription, it was discovered that in addition to RNA polymerase II, the enzyme responsible for mRNA transcription, a large cohort of additional proteins were required to transcribe genes. These ~30 factors are termed General Transcription Factors (GTFs), and are part of the machinery necessary for the production of mRNA in eukaryotic cells (Hampsey 1998). However, since these proteins are ubiquitously expressed, the question of how cell-type specificity of expression was achieved remained open.

A key advance into this question came with the identification of distal regulatory regions called enhancer regions (Banerji et al. 1981). Discovered from SV40 viral DNA, these sequences were shown to drive high levels of transcription regardless of their orientation and at distances several kb from the Transcription Start Sites (TSSs) of genes. Following this, regions upstream of cell-type expressed genes were transfected into different cell lines. Crucially, some of these regions drove expression only in the same cell types as the downstream genes were expressed in. For example, a region upstream of the insulin gene drove expression specifically in a pancreatic cancer cell line (Walker et al. 1983). These sequences where then used to purify the proteins that bound to them, revealing transcriptional activators and repressors.

With this historical background, I will provide the modern conception of transcriptional regulation, particularly in mammals, with a focus on identifying the major players and their means of action.

1.1.1 The role of transcription factors in gene regulation

As mentioned, the primacy of TFs in gene regulation has long been recognized since their identification in the 1960s. Recent research has only solidified their status in establishing cellular states. A good example of the importance of TFs in orchestrating programs of gene expression is demonstrated by the ability of ectopically expressed TFs to reprogram fibroblasts into induced pluripotent stem cells (Takahashi and Yamanaka 2006; Yu et al. 2007).

TFs influence gene expression by binding to *cis*-regulatory elements, typically between 6-20 bp, that are present in the proximal promoter or in distal regulatory regions (Vavouri and Elgar 2005). Once TFs bind to their specific sites, they carry out several functions that have significant consequences for the genes that they regulate. A few of the mechanisms of action include directly recruiting PolII to the TSS (Boehm et al. 2003), promoting the elongation of paused PolII (Kanazawa et al. 2003; Rahl et al. 2010), recruiting chromatin modifiers to the locus (Chen and Dent 2014), facilitating looping of distal regulatory regions to the promoter (Deng et al. 2012) and shuttling loci to distinct regions in the nucleus (Schoenfelder et al. 2010; Kohwi et al. 2013). Through a combination of several of these actions they either activate or repress downstream genes

leading to the elaborate spatial and temporal patterns of gene expression observed in development and response to environmental stimuli.

How TFs find their targets in the genome has been a subject of intense research. Especially in mammalian organisms, across the genome, there are several hundreds of thousands of matches to the motifs that TFs bind to. However, genome localization of these proteins has revealed that they only bind to a fraction of these sites *in vivo*. Further, it is observed that the same TFs can bind to different subsets of TFBSs across different cell types (Lodato et al. 2013).

A major determinant of which TFBSs are bound by a TF is the accessibility of the sites to the protein (Elnitski et al. 2006). Specifically, nucleosomes can act as a barrier to the binding of a TF to a region of the genome. One technique to assay the accessibility of different regions of the genome is DNase-seq (Boyle et al. 2008a). In this assay, nuclei are isolated and then the enzyme DNaseI is used to cut DNA at regions of the genome not bound by nucleosomes. The regions of the genome cut by DNaseI are bound by other proteins, including TFs, general TFs and the PolII initiation complex. Reads from these cuts are then sequenced, aligned to the genome and accessible regions or regions of open chromatin are identified (Song et al. 2011). For example, this technique was applied to identify what determined the binding sites of the Glucocorticoid receptor (GR) (John et al. 2011). The authors used DNase-seq to map the regions of open chromatin in an adenocarcinoma cell line prior to the treatment with a steroid hormone. Upon stimulation with a steroid hormone, the GR is activated and binds to >8000 genomic loci.

Crucially, over 90% of these binding sites are in regions that are accessible prior to stimulation. This implies that accessibility to chromatin is a determinant of which TFBSs among the several hundred thousand across the genome are actually bound by TFs.

This observation then begs the question as to why certain regions are accessible and other regions are closed. One explanatory factor, in addition to the sequence preferences of nucleosomes (Kaplan et al. 2009), is a subset of TFs known as pioneer factors (Zaret and Carroll 2011). These pioneer factors have the unique ability across TFs to bind to their cognate binding sites despite a closed chromatin conformation. Therefore, these TFs bind to their targets even in a closed conformation and open up the neighboring regions allowing other TFs to access their cognate TFBSs. It is important to note that, despite their ability to bind regions that are inaccessible to other TFs, pioneer factors still only bind a subset of all their sites in the genome. This indicates that there are still other factors that regulate the selection of binding sites by pioneer TFs. Interestingly, these factors might themselves be chromatin modifications such as H3K9me3 that have been deposited on specific regions on the genome along the developmental lineage of the cell (Soufi et al. 2012).

1.1.2 The role and identification of distal enhancers in gene regulation

To regulate the expression of genes, TFs typically bind in dense clusters in distal regulatory regions known as enhancers (Bulger and Groudine 2011; Ong and Corces 2011; Buecker and Wysocka 2012). Enhancers have been called the 'information

integration hubs' that consolidate environmental and developmental cues to produce different patterns of gene expression (Buecker and Wysocka 2012). These regions that are typically a few hundred base pairs in length are observed to be, particularly in mammals, at significant distances from the promoter of the gene that they regulate. In some extreme cases, the enhancer can be an Mb away from the gene or even on a different chromosome (Lettice et al. 2003). Recent work in Genome Wide Association Studies linking loci to diseases have revealed that 88% (Hindorff et al. 2009) of disease causing variants lie in non-coding regions of the genome, particularly in enhancers, stressing the importance of identifying and understanding the mechanistic functions of these enhancers (Cowper-Sal lari et al. 2012).

An immediate question that arises is how these distal enhancers influence the transcription of the genes they interact with. Several models have been proposed from observations in different model systems (Ptashne 1986). For example, in the tracking model, regulatory proteins at the enhancer recruit PolII and other initiation factors which then slide along the DNA to find the promoter of the gene they regulate. In more recent literature, the model that has received a preponderance of support is the looping model. In this model, proteins bound at the enhancers interact with those bound at the promoter and thereby influence the expression of the gene. In one example, tethering a co-factor involved in the looping of an enhancer to a gene was sufficient to recruit PolII and to initiate the transcription of the gene (Deng et al. 2012). While this is an important step to understanding the mechanistic functions of enhancers, answers to the key

question of what is mechanistically achieved by the act of looping enhancers to promoters still remain preliminary (Krivega and Dean 2012).

Given the importance of enhancers in regulating gene expression, significant research has gone into methods for identifying these regions. As mentioned above, these regions are a few hundred base pairs and can be several kb away from the genes they regulate. This makes their identification in the large non-coding genomic space problematic (Haeussler and Joly 2011). One attempt to identify these regions, which has been possible with the sequencing of the genomes of several species, has been to look for conserved regions of non-coding sequence (Pennacchio et al. 2006). However, there are two significant drawbacks with this approach. First, it has been noticed that several enhancers which have functional impacts on driving the expression of downstream genes, nonetheless are not conserved (Blow et al. 2010). This implies that one cannot assume that the lack of sequence conservation implies that a region is not functionally relevant in driving expression. Second, even when enhancers are identified by looking at their sequence conservation, there are no clues as to when and in what condition these regions drive gene expression, which makes functional validation of predictions difficult.

Other computational techniques attempt to identify enhancers by looking for sequence similarity across a set of known enhancers that drive a specific pattern of gene expression. These are then used to predict, in a genome-wide fashion, the location of other enhancers that are likely to drive similar patterns of gene expression (Kantorovitz

et al. 2009). For example, these methods have been successful in discovering enhancers that are active in *Drosophila* development. However, this approach is contingent on having a relatively large set of well characterized enhancers for the specific pattern of expression that one is interested in. Further, these methods have been shown to have a far lower efficacy in successfully predicting enhancers in the large non-coding genomic regions such as those found in mammals.

An experimental approach to identifying enhancers is to conduct ChIP-seq assays for TFs that are known to be important in the differentiation of the cell type in question (Chen et al. 2008; Wilson et al. 2010; Tijssen et al. 2011). While individual TFs seem to only weakly predict the regulation of the genes adjacent to which they bind, the clustered binding of TFs has been seen to correlate with active enhancer activity. Crucially, this approach reveals the enhancers that are relevant to the gene expression observed in the specific cell type. Nonetheless, this approach requires one to know the TFs involved, have antibodies that are of ChIP grade quality and sufficient biological material to conduct these assays. In several developmental contexts, these prerequisites cannot be met.

The alternative approach that has gained widespread use across a range of biological systems is to conduct genome-wide profiling to map some combination of chromatin modifiers (e.g. P300) (Visel et al. 2009), mediator subunits (e.g. Med1) (Whyte et al. 2013), chromatin accessibility (DNase-seq or FAIRE-seq) (Xu et al. 2012; McKay and Lieb 2013), chromatin modifications (H3K4me1 or H3K27ac) (Creyghton et al. 2010;

Rada-Iglesias et al. 2011; Rada-Iglesias et al. 2012; Nord et al. 2013; Visel et al. 2013) or enhancer RNAs (Kim et al. 2010; Andersson et al. 2014). Compared to other methods, a significantly larger proportion of enhancers identified by these methods have been shown to accurately drive cell-type specific expression. Importantly, since one only needs to profile a few features correlated with enhancers, this approach also requires less biological material than mapping the location of several TFs. However, it is important to recognize that while these features are highly correlated with enhancers, this does not imply causation (Henikoff and Shilatifard 2011). While factors such as Med1 and P300 are implicated in enhancer function with significant mechanistic basis, the roles, if any, of chromatin modifications still remain to be discovered.

1.1.3 Building predictive models of gene expression

Building predictive models of gene expression has been the focus of significant research. The general approach used is to first identify cell-type specific or temporal patterns of expression that need to be explained. Following this, putative regulatory regions for each gene are identified and scanned for sequence motifs. These sequence features are then used to build predictive models to differentiate the various patterns of expression for different classes of genes. These models are useful for two reasons. First, the performance of these models in predicting different classes of expression allows us to know whether our current understanding of how gene expression is regulated is sufficient to explain the behavior observed. Second, these models can identify specific

sequence features that are predictive. This is particularly useful as it can be used to generate hypotheses regarding the regulators that might bind to these sequence features. These can then be perturbed and gene expression of downstream genes monitored to expedite the process of discovering new regulators of gene expression in the system in question.

Due to the inability to identify distal regulatory regions, early approaches concentrated on using the sequence around the promoter region to predict the expression of genes (Wasserman and Fickett 1998). This approach met with some success in organisms with smaller genomes such as yeast and even organisms such as Drosophila (Beer and Tavazoie 2004; Segal et al. 2008). Early attempts in mammalian organisms with this approach revealed that cell-type specifically expressed genes have lower CG content in their promoters compared to housekeeping genes (Lercher et al. 2003). Further, while there are a few examples where sequence features in promoters can help predict genes that are cell-type specifically expressed (muscle and liver, for example) this approach has met with only limited success.

The key to improving these models likely lies in being able to identify distal regulatory elements and identify the TFs bound in these regions to predict gene expression. In fact, using the ChIP-seq data sets available for a dozen TFs in ES cells, Ouyang et al. showed that not only could they predict relative expression levels, but also absolute expression levels (Ouyang et al. 2009). However, for reasons mentioned above, collecting TF ChIP-seq data for several factors is not feasible in most biological contexts.

As a result, there is a need to combine the approach of building predictive models from sequence features with the genome-wide identification of enhancers using marks correlated with cell-type specific enhancers.

One other approach to build models that has significant predictive power bears mentioning. These models use ChIP-seq data for histone marks around the promoter which are shown to be highly correlated with gene expression patterns and are extremely useful for predicting gene expression (Karlic et al. 2010). However, as mentioned previously, it is important to keep in mind that the correlative or causative impact of histone marks is a question that is yet to be resolved. In fact, some histone marks are a consequence of transcription as opposed to a means to regulate transcription (Henikoff and Shilatifard 2011). Further, in the few cases where the mechanistic consequences of histone marks have been studied, the results have been quite unexpected (Margaritis et al. 2012; Venkatesh et al. 2012). As a result, in the work presented in this thesis, we did not consider comparisons to models of gene expression built from ChIP-seq for histone marks.

1.2 Mammalian sex determination

The differentiation of the gonad is an excellent system to study the principles underlying cell fate determination and the transcriptional programs that drive these processes. Given its central importance to the survival of species, a historical view of how our current conception has evolved is enriching. Therefore, I begin by surveying the key experiments leading to our current view of sex determination. Following this, I discuss the factors involved in the differentiation of the supporting cells in the gonad.

1.2.1 A brief history of sex determination and differentiation

The differences in physiology between males and females of several species have in all likelihood always been noticed. Various explanations have been offered including astronomical signs such as the presence or absence of a halo around the moon during conception, and the 'heat' of the father in determining the sex of the child (Carlson 2013). While examples of this sort are numerous and often entertaining, the modern view of sex determination arose in the early 20th century. The prevailing belief around this time was that sex was determined by external environmental stimuli, such as nutrition, as opposed to internal genetic causes (Morgan 1903).

The first hints that there was a genetic basis of sex determination came from the cytological work of Hermann Henking, Clarence McClung, Nettie Stevens and E. B. Wilson (Brush 1978). These experiments provided important correlative evidence of the presence and absence of specific chromosomes with being male and female. For example, using the mealworm beetle, *Tenebrio Molitor*, Nettie Stevens identified that

male mealworms made sperm with either 10 chromosomes of a large size or 9 large chromosomes and 1 small chromosome. Crucially, when she looked in the somatic cells of males and females, she found that females have 20 large chromosomes, while males have 19 large chromosomes and 1 small chromosome. She concluded that 'this result suggests that there may be in many cases some intrinsic difference affecting sex' (Stevens 1905). The large unpaired chromosome in males was later termed the X and the small chromosome, if present, was termed the Y chromosome. Shortly thereafter, Theophilus Painter conducted careful cytological work in human testicular tissue, discovering that humans too had a difference in their chromosomal complement between males and females (Painter 1923). Specifically, in males, he found one X and one Y chromosomes where females had two X chromosomes.

These discoveries were significant in identifying the potential causes of sex determination in mammals and in other species. However, particularly in mammals, a key question remained regarding the role of the Y chromosome in sex determination. Two chromosomal abnormalities observed in human populations helped settle this question. First, it was shown that patients with Klinefelter syndrome, who display male sexual characteristics, have a 47, XXY karyotype (Jacobs and Strong 1959). Second, it was discovered that patients with Turner syndrome, who display female sexual development, have a 45,X karyotype (Ford et al. 1959). These results strongly confirmed that, in humans, the presence of a Y chromosome was a dominant determinant of male

sexual differentiation and the presence of a single X chromosome was sufficient for female development.

While identifying the importance of the genetic underpinnings of sex determination, a key question remained untouched. Namely, does the organism make cell-autonomous decisions resulting in the various sexual differentiation characteristics or is there a specific organ that is primarily affected by the genetic difference that is then responsible for the sexual differentiation of the rest of the embryo?

The key experiment in determining the causal role of the gonad and the hormones it produces in development was conducted by Alfred Jost in 1947 (Josso 2008). In a technically challenging experiment, Jost removed gonads from rabbits *in utero* and then allowed them to develop. He noted that, while the removal of gonads from female rabbits at any stage had no impact of the development of female characteristics, the removal of male gonads had a time-sensitive phenotype. Specifically, late removal of the gonads from male rabbits (day 23 onward) had no influence on the development of male characteristics. However, removing the gonad at earlier stages in male rabbits caused them to develop as phenotypical females. Later work showed that the two hormones produced by the testis that were instrumental in producing male sexual development were testosterone and Anti-Mullerian Hormone. This was a convincing argument that, in mammals, the gonad was the central organ involved in governing the sexual differentiation of the rest of the organism.

In sum, the modern view of sex determination and differentiation is as follows (Figure 1). Sexual development, in mammals, can be thought of as a hierarchy with three levels. Genetic sex is set at fertilization when an X-bearing egg is fertilized by an X- or Y-bearing sperm. In a process known as primary sex determination, this genetic difference leads to a difference in gonadal sex, i.e. the development of either an ovary in XX embryos, or a testis in XY embryos. The gonads then produce the hormones necessary to direct the sexual differentiation of the rest of the embryo.

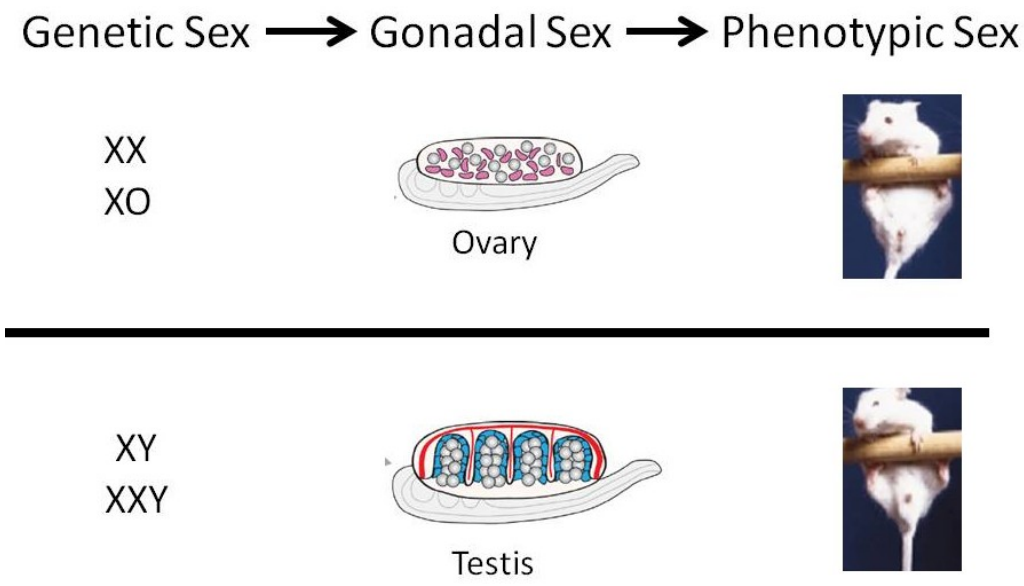


Figure 1: Primary sex determination and secondary sex differentiation

In mice, genetic sex is determined at fertilization by an X or a Y bearing sperm. In a process known as primary sex determination, this genetic difference governs the differentiation of the gonad to an ovary (XX) or a testis (XY). The gonads produce the hormones necessary for the sexually dimorphic differentiation of the rest of the mouse in a process known as secondary sex differentiation. Picture of mice are from (Bishop et al. 2000).

Although the dominant role of the Y-chromosome in determining testis differentiation in mammals was made in 1960, it took 3 decades to identify the gene on the Y chromosome that was responsible for this behavior. After a few missteps, the transcription factor *Sry* was identified as the testis determining factor. The key data came from analyzing three XX individuals who had developed as males. These three individuals shared a 35kb region of the Y-chromosome in which the gene *Sry* was present (Gubbay et al. 1990). In addition, conservation across mammals, testis specific expression of this transcript and the generation of XY female mice confirmed that *Sry* was the causative gene (Koopman et al. 1990; Sinclair et al. 1990) in determining testis development.

1.2.2 Differentiation of the supporting cells: The basis of primary sex determination

In mice, the gonad forms as a thickening of the cells of the coelomic epithelium on the mesonephros around mid-gestation (~ embryonic day (E) 11.0) (Brennan and Capel 2004). Unlike any other organ, the gonad, at this stage, is bipotential - regardless of the genetic sex of the cells that comprise it, it can become a testis or an ovary. Interestingly, even prior to its differentiation at E11.0, it is comprised of multiple cell types including the supporting cells, interstitial cells, endothelial cells and the germ cells. The first cell type to differentiate in XY and XX gonads are the supporting cells. Upon differentiation, these cells provide the signals for the differentiation of the other

cells in the gonad. Therefore, the initial key steps of gonadal differentiation are carried out in the supporting cells.

In murine XY gonads, the TF *Sry*, is transiently expressed between E10.5 and E12.0 and binds to the testis specific enhancer region of another TF on an autosomal chromosome, *Sox9* (Gubbay et al. 1990; Lovell-Badge and Robertson 1990; Vidal et al. 2001; Sekido et al. 2004). The resulting up-regulation of *Sox9* is sufficient for testis differentiation and obviates the need for the expression of *Sry*. It is important to note that there is a critical window for the activation of the male pathway. In a particularly illuminating experiment, Hiramatsu et al. used transgenic mice that had an *Hsp-Sry* construct to initiate the male pathway at specified time points (Hiramatsu et al. 2009). Interestingly, there was a 6hr interval between E11.0 and E11.25 where *Sry* expression had to occur for testis development to be initiated. If this interval was missed, while *Sox9* could be activated, the targets downstream of *Sox9* remained at expression levels observed in the ovary. Indeed, the role of the timing and level of *Sry* activation has been suggested to be involved in sex reversals observed in the case of weak *Sry* alleles.

An important target of *Sox9* is the signaling ligand *Fgf9* which binds to its receptor *Fgfr2* (Kim et al. 2007). This signaling pathway is an important component that reinforces the fate commitment decision by repressing the alternative ovarian pathway and maintaining the expression of *Sox9* (Colvin et al. 2001; Kim et al. 2006). In XX gonads, the signaling ligand *Wnt4* has a central role in directing the ovarian pathway partly through the stabilization of beta-catenin (*Ctnnb1)* (Vainio et al. 1999). Crucially,

stabilization of *Ctnnb1* results in antagonism of the male pathway and the development of XY gonads to ovaries (Maatouk et al. 2008). It is also important to note that the two signaling pathways that promote the testis and ovarian fate are mutually antagonistic (Kim et al. 2006; Kim et al. 2007; Jameson et al. 2012a). Specifically, on a background susceptible to sex reversal, disruption of FGF signaling, results in XY embryos forming ovaries (Colvin et al. 2001). However, when WNT signaling is also inhibited, the phenotype is rescued and XY embryos develop testis (Jameson et al. 2012a). This indicates that the feedback interaction of FGF signaling on *Sox9* expression is likely enacted through the repression of WNT signaling, which itself is repressing *Sox9* expression.

In addition to these genes, genetic experiments have revealed approximately 30 other genes to be involved in the sex determination decision. Specifically, TFs such as *Nr5a1, Wt1, Emx2* and *Lhx9* are involved in the initial development of the bipotential gonad (Eggers and Sinclair 2012). Further, a host of other TFs, co-factors and chromatin modifiers are involved in the differentiation pathways of the testis and ovary (Figure 2). These genes are involved in the dynamic regulation of several hundreds of genes that are differentially expressed between testis and ovaries. Key questions remain regarding how this is accomplished.

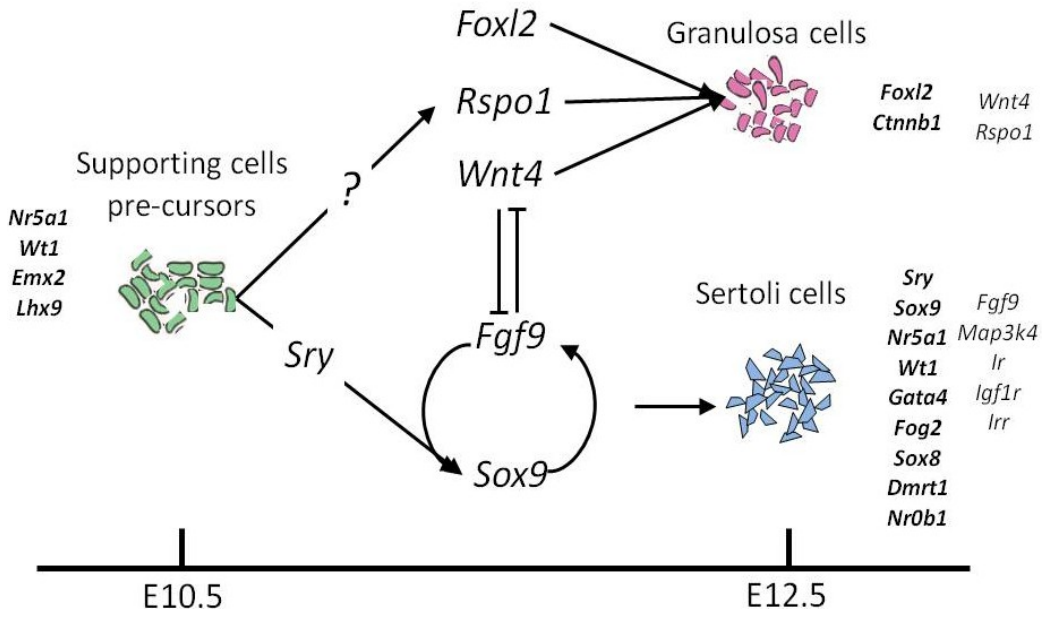


Figure 2: The network in supporting cells

Several genes involved in the differentiation of the bipotential supporting cell pre-cursors to Sertoli cells in the testis and Granulosa cells in the ovary. Transcription factors and co-factors are shown in bold on either side of the figure and signaling molecules are shown on the right of the figure. The basic genetic network driving sex determination is shown in the middle. *Sry* activates *Sox9* which along with *Fgf9* governs the differentiation of the Sertoli cells. While the trigger of ovarian development is as yet unknown, *Wnt4*, *Rspo1* and *Foxl2* are known to be important in governing Granulosa cell differentiation. FGF and WNT signaling also mutually antagonize each other.

Despite identifying several factors involved in primary sex determination (Eggers and Sinclair 2012), several lines of evidence indicate that our knowledge about this network is limited. First, more than half (50-75%) of the human cases of Disorders of Sexual Development (DSDs) are unexplained by mutations in known genes or their regulatory regions. Second, even though we have identified several key regulators, barring a few exceptions (De Santa Barbara et al. 1998; Wilhelm et al. 2007; Bradford et al. 2009), we are unaware of the targets of their regulation. Importantly, gene expression profiling of the gonads done by the Capel lab and others reveal that there are large cohorts of differentially expressed genes soon after the onset of gonadal differentiation (Nef et al. 2005; Munger et al. 2009).

1.3 How to assemble a network of sex determination

To build a predictive model of transcriptional regulation, one can approach the problem as follows. First, the transcriptional behaviors have to be characterized. This would include trying to identify patterns of up- and down-regulation and of expression patterns that do not show variation between conditions. Second, given the importance of TFs in transcriptional regulation, one can use the expression data to identify TFs that also show patterns of differential expression. Third, especially in mammalian organisms, one needs to be able to identify the distal regulatory regions where TFs bind. Finally, the combination of all this information can be used to build predictive models of expression to understand the roles played by the TFs in the expression patterns observed.

This was precisely the approach we took in using data from 19 human cell lines from diverse origins for which we had DNase-seq data identifying putative regulatory regions and matched expression data from Affymetrix microarrays (Natarajan et al. 2012). We then used this information to identify sequence features that predict the expression patterns we observed across these diverse cell lines.

In applying this method to understanding gonadal differentiation, our first step was to carefully analyze the global changes to the transcriptome during the critical window of sex determination (Munger et al. 2013). By observing XX and XY gonads from two different inbred strains of mice, we were able to identify cascades of differential expression that occurred as the bipotential gonad committed to its testicular

or ovarian fate choice. These cascades of expression suggested hypotheses regarding the important TFs in mammalian sex determination.

In order to identify enhancers in a genome-wide fashion in the supporting cells in the gonad, we performed DNase-seq and ChIP-seq for histone marks correlated with active enhancers. We then analyzed these putative regulatory regions to identify the TFs that were bound to these regions.

2. Materials & Methods

2.1 Predicting cell-type specific expression from regions of open chromatin

2.1.1 DNase-seq

DNase-seq was performed on 19 human cell lines representing a wide variety of tissue types, and aligned reads were used to define DNase hypersensitive sites (DHS). Data from 7 cell lines were previously published (Song et al. 2011), and remaining libraries were processed as described in that study. Reads generated were aligned to the hg19 genome using BWA and were then smoothed using a kernel density estimator, f-Seq (Boyle et al. 2008b; Li and Durbin 2009). Following this, DHS peaks were identified as having a -log10(p-value) >= 1.3. We refer to these regions as DHS or regions of open chromatin. Note that the AoSMC cell line was in serum free media.

2.1.2 Classifying DHS based on genomic location

The Refseq hg19 database was downloaded from the UCSC genome browser and used to classify DHS based on their genomic location. If a DHS overlapped the TSS of any transcript variant of a gene it was classified as being a TSS DHS for that gene. Other DHS were similarly classified as Gene Body DHS if they overlapped any region of the gene excluding the TSS. All other DHS were classified as Intergenic DHS.

2.1.3 Microarrays

We used Affymetrix Human Exon 1.0ST microarrays to measure gene expression following ENCODE protocols. We normalized 110 microarrays (measuring 40 cell lines) together, then extracted the subset (19 cell lines) used in the present study. Probesets flagged as cross-hybridizing were first removed from the analysis (Salomonis et al. 2010). Though these arrays provide exon-level probesets, we sought gene-level expression estimates, so we grouped probesets by gene for normalization (Bemmo et al. 2008). To normalize, we used Affymetrix Power Tools (APT) with the chipstream command "rma-bg, med-norm, pm-gcbg, med-polish". This chipstream calls for an RMA normalization with gc-background correction using antigenomic background probes. After normalizing, we noticed an effect due to a switch in microarray reagents. Some of our arrays were processed differently, because our earlier array reagents become unavailable partway through the experiment. Using hierarchical clustering and multidimensional scaling, we found the arrays to group on the basis of reagent used, rather than by biological relatedness. To make the arrays comparable, we used an R script (ComBat) to correct for this batch affect (Johnson et al. 2007). After correction with ComBat, the arrays grouped according to expected biological similarity.

2.1.4 Cell-type specific expression

We identified the genomic location of genes based on matching gene symbols to the refseq hg19 database. If a gene from the array did not have a matching gene symbol,

it was dropped from the analysis. If a gene did not have expression above background (> 4.8) in at least one cell type it was dropped from the analysis. For remaining genes, expression values across the 19 cell lines were z-transformed. The z-scores for expression in each cell line were sorted. The top 200 genes were classified as UR genes, and the bottom 200 genes as DR genes in that cell type. For each cell type, the UR-Other genes were compiled as follows. We first made a set comprising all UR genes from all cell types. We then removed the current cell type UR genes from this global UR gene set. We further removed genes from this set that had expression z-score >= 0 in the current cell type to exclude genes that had higher than mean expression in that cell type. This ensured that this set of genes was purely comprised of genes that were up-regulated in other cell types and had lower than mean expression in the current cell type. A similar procedure identified DR-Other genes. To identify constitutively expressed genes, we selected genes in either UR or DR sets across all cell lines, were expressed above background in all cell lines and had a maximum |z-score| < 1.7 which resulted in 168 genes. This size compared well to the other positive sets of UR and DR genes. This gave us a list of genes that did not have a significant variance in their expression.

2.1.5 GO Analysis

DAVID was used to identify functional categories of genes up-regulated in each cell type (Huang et al. 2009a; Huang et al. 2009b).We used the GO categories and also

the UP_Tissue category to identify the tissue showing the closest gene expression profile to the cell type in question. p-value <0.05 was used as the significance threshold.

2.1.6 PWM scans of DHS and promoter sequences

We collected PWMs for vertebrate TFs from the Transfac, JASPAR and UniProbe databases. This gave us a collection of 789 PWMs some of which refer to the same TF. Note that we allowed multiple PWMs for each TF in our data set since it is generally not known which one reflects binding affinities more accurately. Furthermore, multiple PWMs may also reflect different binding modalities for the same factor. This could be due to, for example, the presence of specific cooperative binding partners in one cell type but not in another.

We scanned the sequence from each DHS site and (proximal) promoter sequence (-900 to +100 relative to each TSS of a gene) using these PWMs. A score was assigned to each location in the sequence based on the log likelihood ratio of the PWM score (probability of the PWM generating the specific sequence) versus the probability that the sequence was generated by a background model. The background model used was a first-order Markov Model trained over a 500bp window centered at the base pair being scored. This effectively corrects for the underlying dinucleotide composition and allows us to separate signal from noise (Megraw et al. 2009). Scores with a log-likelihood ratio less than 0 were not included in further analysis.

After these scores were generated for each base pair, we summed scores across a sliding window of size 60 bp to account for local clusters of multiple, potentially overlapping binding sites. Clusters of binding sites have been shown to be more likely to be bound by TFs as opposed to single binding sites (Gotea et al. 2010). For each DHS or promoter sequence, we assigned the maximum sliding window score as the TFBS score for that TF for that sequence.

2.1.7 Associating DHS TFBS scores with genes

To associate each DHS with a gene that it was potentially regulating, we found the closest TSS to the midpoint of the DHS. The DHS was then associated to that gene. In general, there were several associated DHS for each gene, and these were assumed to be the putative regulatory regions for that gene. To assign one score for a TFBS to each gene, we picked the maximum TFBS score across all the associated DHS (closest gene DHS). In Split DHS, we separated DHS into two groups - overlapping TSSs and those in other parts of the genome. We selected the maximum from both sets, therefore having two features per gene per TF.

2.1.8 Sparse Logistic Regression Classifier

We used a sparse logistic regression classifier that minimizes an objective function which is a linear combination of the sum of squared residuals and the l1-norm of the weights (Koh et al. 2007). We divided our data into 4 parts to perform a 4-fold

cross validation. 3 parts of the divided data were used as the training set. This training set was further divided into 6 parts, 1 of which used as the validation set to learn the hyperparameter for the contribution of the l1-norm in the objective function. The optimal hyperparameter was selected from 10 values (0.001, 0.011 . . . 0.091). This value was then used to evaluate the performance of the model on the original test set, and the area under the receiver operating characteristics (AuROC) was computed as a measure of performance. We performed 5 iterations of each 4-fold cross validation, with the data shuffled before each iteration. Results for each classification task are therefore averages over 20 different partitions of the data, which makes our result more robust to chance arrangements of the data.

2.1.9 Cell-type specific expressed non-redundant TFs

To identify TFs that were cell-type specifically expressed, we used the absolute z-score and extracted gene symbols for TFs with available PWMs. We again used the gene symbol from the Affymetrix arrays to match expression to TF names in our PWM list. By using absolute z-score we picked out genes that were cell-type specifically up- and down-regulated. To ensure that we were picking a non-redundant set of TFs, we used STAMP (Mahony and Benos 2007) with the default settings. Starting from the TF with the highest absolute z-score we only added a TF when it had a significantly different PWM (e-value > 0.25) from the TFs already chosen. We stopped adding to the set when we had 10 TFs.

2.1.10 Conservation analysis

Genomic regions from the 46 way placental mammal track were downloaded from UCSC genome browser. Only regions of size 100bp or more were used. Coding exon sequences were extracted from the refseq hg19 database. BedTools (Quinlan and Hall 2010) was used to subtract exonic coding sequences from conserved regions to find non-coding conserved regions. Regions were scanned for motifs and TFBS scores assigned to genes in the same way as open chromatin regions.

2.1.11 Discriminative Motif finding using MEME

We used MEME to first calculate a position-specific prior using the psp-gen tool. For example, if we wanted to identify the motifs in the DHS sequences related to the UR genes when classifying against DR genes, the positive file was the UR genes and negative file was the DR genes. This was then used to identify motifs from a 1000 positive sequences, UR gene DHS sequences in the above example. The motif width was chosen to be between 8-12 bp and motifs were allowed to have zero or one occurrence per sequence. Motifs with an e-value below 0.05 were then compared to either the top10 non-redundantly expressed TFs or the full list of TFs using STAMP. A STAMP e-value less than 0.05 was accepted as a good match of a motif to TF.

2.1.12 ChIP analysis

ChIP-seq peak coordinates were obtained from ENCODE webpage in BED (narrowPeaks) file format. To assess whether DHS are really bound by the TF or not, we checked for overlap between the coordinates of DHS and ChIP-seq peaks for that TF. To calculate AuROC, the TFBS scores in the DHS were compared against ChIP-seq peaks. Overlaps with the 60 bp window around the TFBS site were considered true positives and others were considered false positives for AuROC calculations.

2.1.13 Footprinting Analysis

DNase reads were used to plot the distribution of DNase-seq reads around transcription factor binding sites. Number of reads mapping to 100 base pairs surrounding transcription factor binding sites were counted for each site and aggregated over the 10,000 highest scoring binding sites. A trough centered at the binding sites in such plots is called a DNase footprint, indicative of protection of the binding site against DNase digestion by a bound TF.

2.2 Fine time-course expression analysis

2.2.1 Mice, dissection, developmental staging, and genotyping

For the time course microarray study, C57BL/6J (stock no. 000664) and 129S1/SvImJ (stock no. 002448) mice were obtained from The Jackson Laboratory. CD-1 outbred mice were used (strain code 022, Charles River) in the gonad primary cell assays.

Timed matings were established for B6 and 129S1, and embryos were collected from dams between embryonic day (E) 11.0-12.0. Embryos were individually staged by counting tail somites (ts) distal to the hindlimbs: E11.0, E11.2, E11.4, E11.6, E11.8, and E12.0 correspond to 13, 15, 17, 19, 21, and 23 ts, respectively (Munger et al. 2009; Michell et al. 2010). For each strain at each time point, three individual pairs of XX and XY gonads from at least two separate litters were collected. The chromosomal sex of each embryo was determined by PCR on head DNA using primers to detect Kdm5c/Kdm5d (5'-TGAAGCTTTTGGCTTTGAG-3' and 5'-CCGCTGCCAAATTCTTTGG-3'). Gonads were dissected away from mesonephroi in sterile PBS (Gibco/Invitrogen, cat no. 1490-144) and stored in RNAlater RNA stabilization solution (Ambion, cat no. AM7024) at -20C until all samples were collected. To minimize contamination and RNA degradation, all surgical instruments and surfaces were treated with RNaseZAP RNase decontamination fluid (Ambion, cat no. AM9780), followed by 70% EtOH in DEPC-treated water, before and during the dissection procedure.

2.2.2 Microarray Processing - RNA isolation, labeling, and hybridization

For the microarray analyses, at least three biological replicate samples were profiled for each strain/stage/sex (n=74 total arrays), with one exception (n=2 replicates for 129S1 E12.0 XY). Total RNA was first extracted from individual pairs of E11.0-E12.0 XX and XY gonads (separated from mesonephroi) with the RNeasy Micro kit with on-column DNase digestion (QIAGEN, cat no. 74004) following the manufacturer's protocol. Total RNA was eluted in 14 ul RNase-free water (not DEPC- treated), and 2 ul were used to quantify RNA concentration on a NanoDrop ND-2000 (Thermo Scientific). Only samples with > 100 ng of total RNA and an A280:A260 ratio of >1.6 were included in the expression analyses.

From each total RNA sample, mRNA was selectively reverse transcribed with oligo(dT) primers to T7-labelled cDNA, and then amplified by in vitro transcription (IVT) to produce biotinylated cRNA using the Illumina TotalPrep Amplification Kit (Ambion/ Life Technologies, cat no. AMIL1791) according to manufacturer's instructions. cRNA concentration was quantified on the NanoDrop ND-2000, and individual samples were concentrated in a vacuum centrifuge as necessary. 750ng of biotinylated cRNA (in ~10 ul volume) were hybridized to Illumina MouseRef-8 v2.0 BeadChips (Illumina, cat no. BD-202-0202) according to Illumina protocols, and array intensity was measured on an iScan scanner (Illumina). To minimize potential for batch effects to confound analysis, individual samples were assigned to 8-sample BeadChips using a balanced design.

2.2.3 Microarray data processing

Microarray data files were imported into GenomeStudio software (Illumina, V2010.1), and raw expression values for each sample extracted. Expression values were quantile normalized and log2 transformed using the R package BeadArray (Ritchie et al. 2011). Probes that had a detection $p < 0.005$ in at least two replicates for any sample type were used for analysis. Data will be publicly accessible on GEO upon acceptance of the manuscript. Data are publically accessible in GEO (accession number GSE41948).

2.2.4 ANOVA analysis

The ANOVA analysis was conducted using the R package Limma (Smyth 2004). A sex by strain by stage factorial analysis was conducted as outlined in (Smyth et al. 2005). The model included the sex, strain, and stage variables, the sex*strain, sex*stage and strain*stage two-way interaction terms, and a three-way interaction term sex*strain*stage. The model was fit for all the probes that had reliable expression (detection $p < 0.005$) in at least two replicates of any one sample using the lmFit function in the Limma package. The statistical significance of each of the terms was evaluated using the eBayes function in Limma. Probes that did not have a significant difference (Benjamini-Hochberg adjusted $p < 0.05$) for at least one of the variables were excluded from further analysis.

2.2.5 HMM to identify dimorphic expression

Hidden Markov Models (HMMs) are generative probabilistic models that explicitly model the observed data as being emitted by a 'hidden' biological state (here, male or female enrichment). Further, transition probabilities between states capture the time dependencies in data between adjacent time points. Inference algorithms allow for computing the most probable state paths that give rise to the observed data, and accounts for noise inherent in observed data. The modeling of time dependencies between biological states, and accounting for noisy observations, makes HMMs particularly well-suited to analyze time course microarray data (Schliep et al. 2003; Yuan and Kendziorski 2006).

We designed a left-to-right HMM with three states per time point. The three states correspond to male state (with higher expression in males), female state (with higher expression in females) and similar expression state (with no difference in expression between the two sexes). The observed data on which the model was trained and clustered was the quantized Fold Difference (FD) of the log2 normalized values between XX and XY gonads at each time point. Note that a fold change of expression of 1.25 corresponds to an FD of 0.3219, fold change of 1.5 to an FD of 0.585 and a fold change of 2 to an FD of 1. Limma was used to calculate FD between XX and XY gonads at each time point for each strain. If a specific comparison did not have a p-value <0.05, or $|FD| > 0.3219$ then the FD for that comparison was set to 0. The FD was then quantized into symbols as follows:

Symbol s - Similar expression in XX and XY gonads [-0.3219 < FD < 0.3219].

Symbol m_1 - Higher expression in XY gonads [0.3219 < FD < 0.5850].

Symbol m_2 - Higher expression in XY gonads [0.5850 < FD < 1].

Symbol m_3 - Higher expression in XY gonads [1 < FD]

Symbol f_1 - Higher expression in XX gonads [0.3219 > FD > -0.5850].

Symbol f_2 - Higher expression in XX gonads [0.5850 > FD > - 1].

Symbol f_3 - Higher expression in XX gonads [1 > FD].

The symbols m_1, m_2, m_3, and f_1, f_2, f_3 indicate varying levels of confidence in the differential expression between XX and XY gonads.

For each gene, for each strain there were 6 symbols indicating the FD between XX and XY gonads across the time window. For example, for Sox9 in the 129S1 strain, the following FDs were observed at the 6 time points – 0, 0.77, 1.41, 2.41, 2.30, and 1.97. Following the rules listed above, this was quantized as s, m_2, m_3, m_3, m_3, m_3.

The emission probabilities of the HMM were initialized as shown in Figure S2B to reflect the intuitive meaning of the states and the possible observed symbols from each state. Note that all probabilities were initialized as being non-zero. After training was completed, emission probabilities still reflected the intuitive meaning of the states (Figure S2B).

The transition probabilities between states were initialized as follows (Figure S2A). Observed symbols were first classified into male, female and similarly expressed – symbols m_1, m_2, m_3 into state M, symbols f_1, f_2, f_3 into state F, and symbol s to state S.

Transitions between all combinations of states in adjacent time points were counted and normalized to make transition probabilities from each node sum to 1. A pseudocount of 1 was added to all possible transitions (transition between states in adjacent time points) to initialize all probabilities as non-zero.

The HMM was trained using the Baum-Welch algorithm for 200 iterations with data from both strains for all the probes that passed the filtering criteria and were shown to have a significant effect for at least one variable in the ANOVA analysis. The state path for the observed FDs for each of the probes was computed using the Viterbi algorithm. Probes with the same state paths were clustered together.

2.2.6 shRNA clones and lentivirus production

Pre-validated gene-specific MISSION® shRNA clones (Sigma Aldrich; Sox9 pLKO.1 clones: TRCN0000086165, TRCN0000086167; Lmo4 pLKO.1 clones: TRCN0000084373, TRCN0000084375; Non-targeting Controls – TurboGFP shRNA SHC004, eGFP shRNA SHC005) and lentiviral packaging and envelope plasmids (Addgene; pCMV-dR8.2 dvpr ID# 8455, pMD2.G ID# 12259) were purchased as bacterial stocks, and high quality plasmid DNA was isolated from overnight liquid LB cultures with a Maxiprep kit (QIAGEN, cat no. 12162) following manufacturer's instructions and quantified on a NanoDrop ND-2000.

Lentivirus production followed the Addgene 4-day protocol with slight modifications (www.addgene.org/tools/protocols/pLKO/) (Moffat et al. 2006). All work

with lentiviruses was performed in a BSL2+ hood following approved biosafety procedures. On Day 1, for each sample, $5x10^6$ HEK-293T/17 cells (ATCC cat no. CRL-11268) were suspended in 10 ml of Dulbecco's Modified Eagle Medium (DMEM, Gibco cat no. 11995) + 10% Fetal Bovine Serum (FBS) without antibiotics, plated to 10 cm cell culture plates, and incubated at 37°C, 5% CO_2 overnight. Late in the afternoon of Day 2, 10 ug of pLKO.1 shRNA plasmid, 7.5 ug of pCMV-dR8.74 dvpr packaging plasmid, and 2.5 ug of pMD2.G envelope plasmid DNA were suspended in Opti-mem serum-free medium with 60 ul X-tremeGENE HP DNA transfection reagent (Roche, cat no. 06 366 236 001) in a 3:1 ratio to a total volume of 600 ul, incubated at 25°C for 20 minutes, then applied drop-wise to the 10 cm plate containing HEK-293T/17 cells at 60-80% confluency, swirled gently to disperse evenly but not dislodge cells from the plate, and incubated at 37°C, 5% CO_2 overnight (12-18 hours). On day 3, media containing the transfection reagent was removed carefully and decontaminated in >10% bleach. Next, 5.5 ml of fresh viral growth medium (vGM, containing Neurobasal medium (Gibco, cat no. 21103-049) supplemented with 10% FBS, 0.5 mM L-glutamine (Gibco, cat no. 25030-149), and 1x Antibiotic-Antimycotic (Gibco cat no. 15240-062)) was added carefully to the side of the plate so as not to disturb the transfected virus-producing cells, and incubated at 37°C, 5% CO_2 overnight. Late in the afternoon of day 4, the virus-containing vGM was harvested with a 10 ml syringe, and filtered through a 0.45 um PES syringe filter (Whatman, cat no. 6780-2504) into sterile 2.0 ml polypropylene cryo-vials. Viral media was stored at 4°C for use within 5 days, or at -80°C for long-term storage. All

laboratory materials that came into contact with viral particles were treated as biohazardous waste and autoclaved according to BSL2+ safety practices.

2.2.7 Gonad Primary Cell Assays

The effect of silencing candidate regulatory genes was assayed in dissociated gonad primary cell cultures. Timed matings were established for CD-1 mice, and embryos were collected from dams at E12.5. Gonads were dissected away from the attached mesonephroi, sexed by visual inspection for testis cords, and XY gonads from a litter were counted and pooled. Pooled XY gonads were then dissociated in 0.25% Trypsin-EDTA (1x, Gibco cat no. 25200-056) for 15 minutes at 37°C with slight agitation, followed by centrifugation at 4000 rpm for 5 minutes, washed once with DMEM (Gibco cat no. 11965) followed by centrifugation, and suspended in Opti-Mem (Gibco cat no. 11058-021) supplemented with 1% FBS. Cells from one pair of XY gonads were determined to be sufficient for one well of a 24-well culture plate, and the amount of suspension liquid was calculated by multiplying 250 ul by the number of pairs of XY gonads in the pooled sample.

Following the dissociation and wash steps, 250 ul gonad primary cells were immediately added to individual wells of a 24-well cell culture plate, and 250 ul of the appropriate lentivirus-containing vGM was added to each well in a BSL2+ hood. In addition to wells designated to assay target gene shRNA-mediated knockdown, separate wells containing XY gonad primary cells from the same litter were infected with the

non-targeting eGFP shRNA (SHC005) and/or TurboGFP shRNA (SHC004) controls. Plates were incubated at 37°C 5% CO_2 for 68-72 hours and cell viability was monitored daily with a light microscope. Virus production could be monitored visually for the TurboGFP control infected cells using a fluorescence microscope.

2.2.8 qRT-PCR

Following 68-72 hours incubation, total RNA was isolated from shRNA lentivirus-infected gonad primary cells using Trizol reagent (Life Technologies, cat no. 15596-018). Briefly, lentivirus-containing culture media was first removed from each well and disposed in bleach. Next, 400 ul of Trizol was added to the adherent cells in each well, allowed to sit at room temperature for 3-5 minutes, after which the lysate was transferred to 1.5 ml microcentrifuge tubes. Subsequent RNA isolation steps follow (Munger et al. 2009). Total RNA was quantified on a NanoDrop ND-2000, treated with DNaseI (Life Technologies, cat no. 18068-015), and converted to cDNA using the iScript cDNA synthesis kit (Bio-Rad, cat no. 170-8891) following manufacturer's instructions.

Gene expression was quantified by quantitative RT-PCR (qRT-PCR) on a StepOnePlus Real-time PCR System (Life Technologies). For qRT-PCR, each analysis was performed in technical triplicate in a total volume of 20 ul reaction mix containing 2 ul cDNA template, 4 ul 1 uM gene-specific forward and reverse primers, 10 ul 2x Quantace SensiMix SYBR (Bioline, cat no. QT615-02), and 4 ul RNase-free water. The list of qRT-PCR primers can be found in Table S3; most have been previously published

(Munger et al. 2009; Jameson et al. 2012a). All primer sets were tested for efficiency and found to work optimally with the ΔC_t method (Simon 2003). Within a sample, target gene C_t thresholds value were determined and normalized to Gapdh. Differences between target gene shRNA and non-targeting control shRNA samples were compared using the $\Delta\Delta C_t$ method as described previously (Simon 2003). Significance of expression differences between samples was assessed using a t-test.

2.3 Identifying active enhancers in XY supporting cells

2.3.1 Collecting cells

Transgenic *Sox9-ECFP* mice (Kim et al. 2007) and *Sox9-ECFP; Sry-EGFP* mice were maintained on a C57Bl/6 background. To isolate pre-Sertoli cells, homozygous males of both genotypes were bred to CD-1 (Charles River) females in timed matings to generate E15.5 or E13.5 embryos. Noon of the day a vaginal plug is observed is defined as E0.5. Embryos were dissected and testes removed from the adjacent mesonephros. Testes from one or more litters were pooled together, incubated in 500 μl of 0.25% Trypsin-EDTA (Gibco) plus 0.25% Collagenase at 37°C of 8-10 minutes. The Trypsin-EDTA was removed and the tissue rinsed with 1X PBS with 3% BSA, and then dissociated in 500 μl 1X PBS with 3% BSA. Dissociated cells were passed through a cell strainer (BD Falcon) to ensure a single cell suspension. Cells were sorted as described previously (Jameson et al. 2012b). For DNase-seq, sorted cells were pelleted, and then resuspended in 250 μl of Recovery-Cell Culture Freezing Media (Gibco) and slowly frozen to -80°C. For ChIP-seq, pelleted cells were resuspended in 360 μl of PBS and cross-linked with 10 μl of 37% formaldehyde at room temperature for 10 minutes. Cross-linking was stopped by addition of 46.3 μl of 1M glycine for 5 minutes at room temperature. Cells were then pelleted, supernatant removed and stored at -80°C.

Mouse tissues (kidney, liver, heart and brain) were collected from adult C57Bl/6 mice, flash frozen and then pulverized before use. The mouse fibroblast cell line was derived from adult C57Bl/6 mice (Jackson Labs). ESCs, also of the C57Bl/6 strain, were

kindly provided by Ute Hochgeschwender (Duke University) and were grown on gelatinized plates in the absence of a feeder layer or matrigel. To harvest ESCs, plates were washed with 1X PBS and treated with 0.25% Trypsin-EDTA for 7-10 minutes at 37°C. An equal volume of medium (containing 10% FBS) was then added to the plates to stop trypsinization. Cells were collected and pipetted up/down to get single cell suspension and were centrifuged at ~1.2 krpm for 10 minutes. All medium was removed from cell pellet.

2.3.2 ChIP-qPCR and ChIP-seq

30 μl of dynabeads Protein A (Life 10002D) were washed twice with 500 μl ChIP Dilution Buffer (CDB) (1 % Triton X-100 (Sigma T8787), 2mM EDTA, 150mM NaCl, 20mM Tris (pH 8)) and resuspended in 30 μl of CDB. Tubes were split equally and 470 μl of CDB with 1x Proteinase inhibitor cocktail (PIC) (Sigma P8340) and 1x Phenylmethanesulfonyl fluoride (PMSF) (Sigma 93482) was added to each tube. Antibodies (H3K4me1 (4 μl) – Active Motif 39298, H3K27ac (2.5 μg) - abcam ab4729) were added to each tube. Tubes were put into an end-over-end rotator for >4 hours at 4°C.

1M cells were pooled from multiple sorts (for gonad cells) or from cultures (C2C12) and washed twice in 500 μl of PBS with 1x PIC and 1x PMSF. Cells were resuspended in 500 μl of lysis buffer (50mM Tris-HCL, 10mM EDTA, 1% SDS) with 1x

PIC and 1x PMSF and sonicated with Branson 450 Sonicator at output power of 3, duty cycle of 30% for 16 cycles of 30 seconds with 1 minute rest time between sonications.

Lysate was spun down at 4°C for 10 minutes at 100krpm. 440 μl of lysate was put into a new tube, with 200 μl added to each pre-incubated dynabeads and antibody (K4me1 and K27ac) tube and 40 μl in a tube for input. We added 700 μl of CDB with 1x PIC and 1x PMSF to IP tubes and incubated in an end-over-end rotator at 4°C overnight. 160 μl of CDB and 8 μl of 5M NaCl were added to the input and the tube was placed at 65°C.

The following day, washes were performed on the IP tubes as follows: Once with Wash Buffer 1 (50mM Tris HCl, 1mM EDTA, 150 mM NaCl, 0.1% SDS, 0.1% Triton X-100, 0.1% Sodium deoxycholate), twice with Wash Buffer 2 (50mM Tris HCl, 1mM EDTA, 500 mM NaCl, 0.1% SDS, 0.1% Triton X-100, 0.1% Sodium deoxycholate), once with Wash Buffer 3 (10mM Tris HCl, 1mM EDTA, 1% NP-40, 1% Sodium deoxycholate, 250mM LiCl), twice with Wash Buffer 4 (50mM Tris HCl, 1mM EDTA, 500 mM LiCl, 1% NP-40, 0.7% Sodium deoxycholate), twice with TE buffer [8.0]. All washes were done with 1mL at 4°C for 5 minutes. 1x PIC and PMSF were added to all buffers. DNA-protein complexes were eluted twice from the beads with 100 μl elution buffer (100mM sodium bicarbonate, 1% SDS, 8mM NaOH). 8 μl of NaCl was added to the eluates and tubes were placed on 65°C overnight. Solutions for wash buffers were mainly gleaned from the protocols posted on the Epigenomics Roadmap website.

All tubes were removed from 65°C heat block and left to cool to room temperature. 1 μl of RNase-cocktail (Life AM2286) was added to each tube. Tubes were incubated for 30 minutes at 37°C. Following this, 4 μl of 0.5M EDTA, 8 μl of 1M Tris and 1 μl of Proteinase K was added. Tubes were incubated at 45°C for 60 minutes. DNA was purified using QIAgen PCR purification columns (28104).

For ChIP-qPCR, qPCR was performed as before. The percent input method was used to calculate enrichment (Haring et al. 2007). For library preparation for sequencing, DNA was concentrated using a vacuum centrifuge to ~10 μl. 10 μl of IP DNA and 1ng of input DNA was used in the library preparation. The Rubicon ThruPLEX FD kit was used for library preparation according to the manufacturer's protocol. Size selection of smaller size amplified DNA was done with SPRI beads (Agencourt AMPure XP A63880) at 0.6x concentration.

ChIP-seq analysis was performed with SICER (Zang et al. 2009). Enrichment was called for the histone modifications using the input track as the control. The species variable was set to mm9, redundancy threshold to 2, window size to 200, fragment size to 150, effective genome fraction to 0.7, gap size to 600 and FDR to 0.01.

2.3.3 Overlap analysis with different regions

Various types of genomic locations were put into bed files including TSS, genic, intronic and exonic. These regions were extracted from UCSC genome browser. Bed files for date from other cell types were used. Overlap was calculated by using bedtools

(Quinlan and Hall 2010) with an overlap of 1bp being sufficient. Similar analysis was done to identify overlap between DHSs and K27ac peaks.

2.3.4 Assigning DHSs to different genes

DHSs were assigned to genes based on their location in an approach similar to the GREAT algorithm (McLean et al. 2010). The domain for each gene was extended 1Mb from the TSS and Transcription End Site (TES) until they encountered the TSS -5kb of another gene. Based on gene categories from (Jameson et al. 2012b), boxplots of the DHSs belonging to these regions were constructed (Figure 34). A two-sided Mann-Whitney test was used to compare the distributions of number of DHSs.

2.3.5 Transient transgenics, Immunocytochemistry and Imaging

A putative regulatory region (UCSC mm9 coordinates chr2:104914099-104915125) upstream of *Wt1* was amplified by PCR and cloned into the NotI site of the *Hsp68 –LacZ* reporter vector (obtained from Addgene; Plasmid #33351). Cloning was carried out using the In-Fusion HD (Clonetech) reagent. To prepare DNA for zygote injection, 50 μg of the *TgWt1* plasmid was linearized with NotI-HF and HindIII and gel purified by electroelution. DNA was phenol-choloform extracted, ethanol precipitated and resuspended in EmbryoMax Injection Buffer (Millipore, MR-095-10F). The DNA was further purified on a DNA-cleanup column (Qiagen) and eluted again in

EmbryoMax Injection Buffer. Pronuclear injections into B6SJLF1/J zygotes were performed by the Duke Transgenic Core Facility to generate transient transgenics.

Embryos were dissected at E13.5 and the embryonic tail was removed for genotyping to detect the *LacZ* gene. Gonads were carefully dissected from embryos and fixed in 4% paraformaldehyde for several hours or overnight at 4°C. The remaining embryo bodies were fixed in 4% paraformaldehyde for 8 minutes, washed in X-gal wash buffer (2mM $MgCl_2$, 0.2% NP-40 in 1X PBS) and then incubated overnight at 37°C in X-gal staining solution (5mM potassium ferrocyanide, 5mM potassium ferricyanide, 1mg/ml X-gal in X-gal wash buffer).

For immunostaining, fixed gonads were washed three times in 1X PBS and incubated in blocking solution (10% FBS, 3% BSA and 0.1% Triton-X-100 in 1X PBS) for 1 hr at room temperature. Blocking solution was replaced with primary antibodies diluted in blocking solution and incubated overnight at 4°C. The next morning, samples were washed three times in washing solution (1% FBS, 3% BSA and 0.1% Triton-X-100 in 1X PBS) followed by one hour incubation with blocking solution. Samples were then incubated with secondary antibodies, diluted in blocking solution, overnight at 4°C. Following three washes, samples were then mounted in DABCO (2.5% 1,4, diazagicyclo octane, 90% glycerol in 1X PBS). Images were taken on a Leica SP2 confocal microscope.

Primary and secondary antibodies were used at the following dilutions: rat-anti-CDH1, 1:250 (Zymed, 13-1900); rabbit-anti-⍵-galactosidase, 1:10,000 (MP Biomedicals, 55976); goat-anti-MIS/AMH, 1:250 (Santa Cruz, sc-6886); Alexa Fluor 488-anti-rat, 1:500

(Molecular Probes, A21208); Cy3-anti-rabbit, 1:500 (Jackson ImmunoResearch Laboratories, 711-165-15); Alexa Fluor 647-anti-goat, 1:500 (Molecular Probes, A21447).

3. Predicting cell-type specific expression from regions of open chromatin

3.1 Summary

Complex patterns of cell-type specific gene expression are thought to be achieved by combinatorial binding of transcription factors (TFs) to sequence elements in regulatory regions. Predicting cell-type specific expression in mammals has been hindered by the oftentimes unknown location of distal regulatory regions. To alleviate this bottleneck, we used DNase-seq data from 19 diverse cell types to identify proximal and distal regulatory elements at genome-wide scale. Matched expression data allowed us to separate genes into classes of cell-type specific up-regulated, down-regulated, and constitutively expressed genes. CG dinucleotide content and DNA accessibility in the promoters of these three classes displayed substantial differences, highlighting the importance of including these aspects into modeling gene expression. We associated DNaseI Hypersensitive Sites (DHS) with genes, and trained classifiers for different expression patterns. TF sequence motif matches in DHS provided a strong performance improvement in predicting gene expression over the typical baseline approach of using proximal promoter sequences. In particular, we achieved competitive performance when discriminating up-regulated genes from different cell types or genes up- and down-regulated under the same conditions. We identified previously known and new candidate cell-type specific regulators. The models generated testable predictions of activating or repressive functions of regulators. DNaseI footprints for these regulators

were indicative of their direct binding to DNA. In summary, we successfully used information of open chromatin obtained by a single assay, DNase-seq, to address the problem of predicting cell-type specific gene expression in mammalian organisms directly from regulatory sequence.

3.2 Introduction

Genome-wide techniques, such as chromatin immunoprecipitation followed by microarrays or sequencing (ChIP-chip and ChIP-seq), have been instrumental in identifying precise TFBS which can then be used to predict gene expression. For example, ChIP data for 12 key TFs in embryonic stem (ES) cells was used to predict both absolute and relative expression values with high accuracy (Chen et al. 2008; Ouyang et al. 2009). While impressive, it is important to note the difficulty in procuring this kind of data across a wide variety of cell types. First, in order to conduct ChIP one needs a high quality antibody or tagged protein, which is not always available for the TF(s) of interest. Second, TFs have to be assayed individually, which requires many independent ChIP experiments to identify combinatorial patterns of TF binding. Finally, for this method to succeed, one must have a good understanding of the cell type in question to know which TFs to analyze. As a result, for most cell types there is not enough information available on the binding profiles of TFs to predict cell-type specific gene expression. Therefore, developing predictive models of gene expression without relying on ChIP would facilitate our understanding of transcriptional regulation.

A more widely applicable alternative to ChIP is to use known cognate binding preferences for TFs determined from assays such as SELEX, ChIP-seq, ChIP-chip and Protein Binding Microarrays (PBMs) (Stormo and Zhao 2010) to find TFBSs in putative regulatory regions. However, without knowing the location of distal regulatory regions, most studies using this method focus exclusively on TFBS identified in proximal

promoter sequences (Das et al. 2006; Ramsey et al. 2008; Sinha et al. 2008; Suzuki et al. 2009). Using these sequence features has revealed, for example, a crucial CG content difference between cell-type specific and constitutively expressed genes in mammalian organisms (Yamashita et al. 2005; Carninci et al. 2006). However, these approaches have frequently struggled to distinguish between more specific patterns, such as predicting cell-type specific expression across many cell types. A comprehensive understanding of cell-type specific expression will require identification of both proximal promoter and distal regulatory elements. While comparative genomics has been successfully used to pinpoint functionally relevant regions, recent reports have stressed the complexity of evolution in functional non-coding regions and the resulting frequent lack of sequence conservation (Ludwig et al. 2005; Odom et al. 2007; Blow et al. 2010).

For over three decades, mapping DNaseI hypersensitive sites (DHSs) has been employed to identify the location of many types of active gene regulatory elements (Wu and Gilbert 1981). DNaseI is an enzyme that preferentially digests DNA in regions of low nucleosome occupancy, i.e. regions of open or accessible chromatin. DHSs have been found to be well correlated with genomic features such as transcription start sites (TSSs), distal enhancers, insulators, transcription factor binding sites, and active histone marks (Heintzman et al. 2007; Boyle et al. 2008a; Heintzman et al. 2009). A recent study profiling open chromatin in 7 cell types in a genome-wide fashion using DNase-seq highlighted that open chromatin regions are similar across functionally related cell types, that cell-type specific regions are distal to TSSs, and identified groups of DHSs

that show coordinated nucleosome depletion (Song et al. 2011). Other studies have indicated that DNase-seq data can be used to identify TFBSs at single-nucleotide resolution (Hesselberth et al. 2009; Boyle et al. 2011; Pique-Regi et al. 2011).

In this study, we use DNase-seq data across 19 diverse human cell lines to define proximal and distal regulatory regions, and to quantify the contribution of sequence features in DHSs to specify different patterns of cell-type specific gene expression. Using expression data from the same 19 cell types, we define classes of up-regulated, down-regulated and constitutively expressed genes, which show distinct patterns of chromatin accessibility. We then build predictive models specifically for these different expression classes, by using the binding site matches that map within DHSs. Crucially, these models dramatically improve on baseline models of proximal promoter regions, and specifically control for the impact of promoter CG content on classifier performance.

Our results demonstrate the crucial role for sequence features in open chromatin regions for determining expression patterns, and its usefulness for building predictive regulatory models. We confirm many known regulatory interactions and identify novel putative positive and negative regulators of gene expression. We also reveal the presence of DNase footprints for specific TFs that are identified as predictive in our model indicating direct binding to DNA. Our work provides a general and easily extensible framework to address questions related to gene regulation in vertebrates.

3.3 Results

3.3.1 DHSs have different properties depending on their genomic location

As part of the ENCODE project, DNase-seq has been performed in several human cell lines representing a wide variety of tissue types. Aligned reads were used to define DNaseI hypersensitive regions (DHS; see Methods for details). Of these, we selected 19 cell lines to represent a broad and largely unrelated variety of cell types. These include DNase-seq data from a recent study across 7 cell lines (Song et al. 2011). In each of the 19 cell lines we used, DHS regions cover ~2% of the genome (Table 1). This indicates that a large proportion of *cis*-elements likely to be involved in establishing the expression patterns in each cell line only comprise a small fraction of the genome. Such regions may encode specific activation patterns of genes, but also include insulators that can define target relationships. A hallmark of insulators is the presence of binding sites for the CCCTC binding factor (*CTCF*). Across the 9 cell types for which *CTCF* ChIP-seq data is available, ~28% of DHS overlapped CTCF bound sites, in agreement with recent work (Song et al. 2011).

Table 1: Number of DHS per cell line and the total number of bases covered by the DHS peaks

Cell Line	Number of DHS	Number of bases in DHS (% of genome size)
Chorion	140042	3.55
Medulloblastoma	151858	2.44
FB0167P	126789	1.91
GM12878	124321	1.96
H1_ES	129061	2.83
Glioblastoma	118222	1.71
Hela S3	141165	2.04
Hepatocytes	171897	3.48
HepG2	116018	1.72
HMEC	158764	2.37
HUVEC	126695	2.08
K562	133372	1.87
LnCAP	144070	2.82
MCF7	119828	1.58
Melanocyte	130073	1.50
HSMMTube	161083	2.68
NHEK	140520	1.80
Osteoblast	151381	2.45
AoSMC	121731	1.62

Based on their genomic location, DHS were divided into exclusive classes as follows. We first identified a set of TSS DHS as those that overlapped the transcription start sites (TSS) of genes based on Refseq hg19 annotation (Figure 3A). Other DHS were designated as Gene Body DHS if they overlapped exons or introns, and as Intergenic DHS if they did not overlap any genes. The median size of all DHS was approximately 300bp, with the TSS DHS set as outlier with a median size of ~ 1kb (Figure 3B). The larger size of TSS DHS may reflect the presence of larger and more stable complexes such as the Pre-Initiation Complex (PIC) near the TSS of genes.

The normalized CG dinucleotide content of Gene Body and Intergenic DHS showed a median of 0.28 and 0.26, respectively (Figure 3C). For TSS DHS, the normalized CG content showed a unimodal distribution with its mode at ~0.8, with a heavy tail of several DHS with CG content below 0.6.

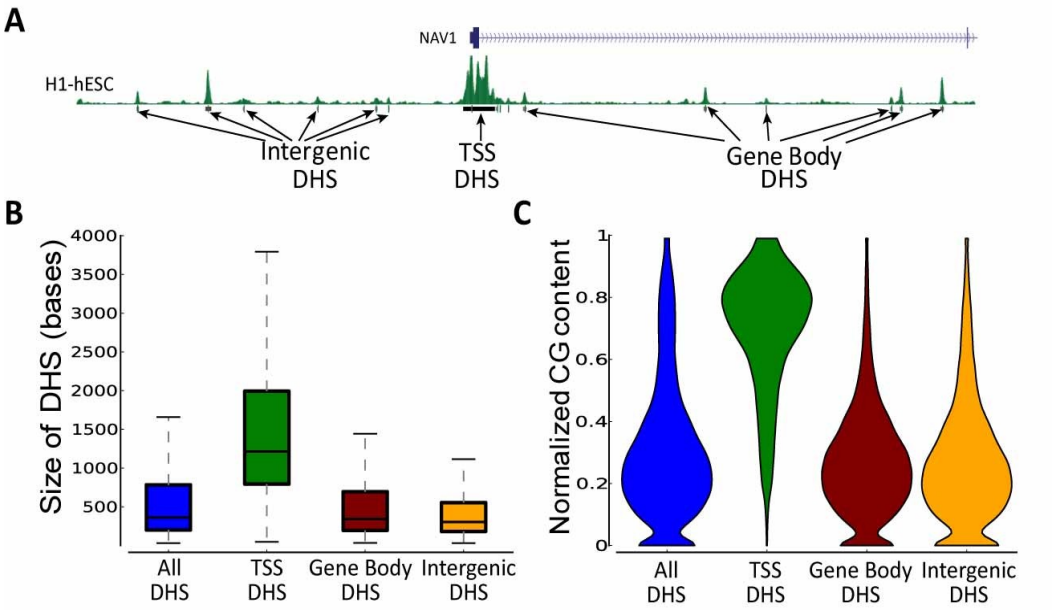


Figure 3: Properties of DHS based on genomic location

(A) DHS that are intergenic and those that are overlapping the TSS and gene body were classified as Intergenic, TSS and gene body DHS, respectively (Chr1: 201,566,484 – 201,683,121). **(B)** Sizes of different DHS for the Chorion cell line. Data from only one cell line was used to avoid multiple counting of ubiquitous DHS. Other cell lines show similar trends. Outliers are not plotted. **(C)** Violin plot showing normalized CG content for different DHS in the Chorion cell line. The subset of DHSs with normalized CG content of zero are comparatively small (median of 128 bp).

3.3.2 A large proportion of TSSs are found in regions of accessible chromatin

To understand how regions of open chromatin vary between cell types, we inspected the degree to which DHSs were shared in the 19 cell types. DHSs were classified as being specific to a cell line if it was only present in a single cell type or overlapped less than 50% of its length with a DHS from any of the other 18 cell types (Figure 4A). Across all DHSs ~14% were specific to a single cell line (Figure 4B). Intergenic DHS showed the highest percentage of being cell-type specific (~17%). Conversely, TSS DHS were largely not cell-type specific with less than 1% being open specifically in a single cell type. Despite the broad panel of cell lines that vary in expression, the chromatin state at the TSS of these genes was open and largely invariant across multiple cell lines. This is in agreement with a recent study analyzing a subset of the cell types used here (Song et al. 2011).

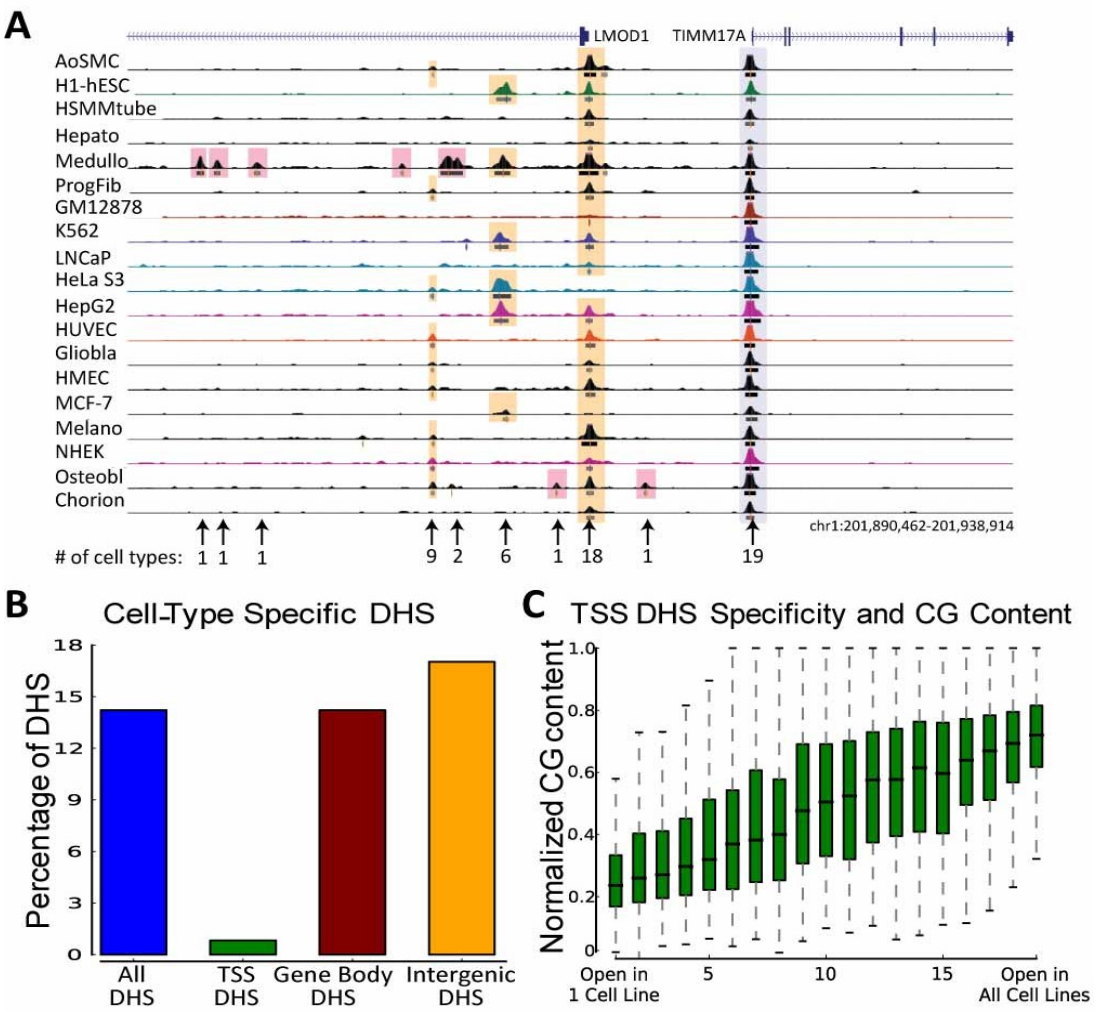


Figure 4: Cell-type specificity of hypersensitive regions.

(A) Example (Chr1: 201,890,462 – 201,938,914) showing cell-type specific DHS across two cell lines. Note that we called a DHS cell-type specific if it did not overlap another DHS by more than half in any of the 18 other cell line. **(B)** Bar graph showing the proportions of cell-type specific DHS across different genomic locations averaged across all cell lines. **(C)** TSSs were divided by the number of cell lines they were in a region of open chromatin. For each set of TSS, normalized CG content in the promoter regions (-900,100) of the TSS are shown.

We determined the normalized CG content in the proximal promoter region of the gene, defining the proximal promoter as -900 to +100 bp around the TSS. If a gene had multiple TSSs, the average of the normalized CG content from each TSS was used. There was a steady positive trend in the number of cell lines in which a DHS overlapped a TSS and the CG content around the TSS (Figure 4C). Previous studies have reported that gene expression can be predicted from the CG content in the proximal promoter region (Yamashita et al. 2005; Carninci et al. 2006; Zhu et al. 2008). Our result indicates that higher levels of CG dinucleotide content, and thus more frequent presence of CpG islands, are positively correlated with, and could be functioning to preserve, an open chromatin state surrounding the TSS. There were fewer genes with a TSS open in only one cell type (976 genes) and many with an open TSS across all 19 cell types (8393).

3.3.3 Cell-type specific expressed genes show differing patterns of accessible chromatin at their TSS

Gene expression data for the 19 cell lines were generated using Affymetrix exon arrays. Expression values for each gene were transformed to z-scores across all the cell lines. Genes with large positive or negative z-score values thus showed a larger deviation from the mean expression across cell types. The z-score transformed expression values were used to select subsets of genes with specific expression patterns (Figure 5A-C). Up-regulated genes, exemplified by *GCM1* (Figure 5A) had a particularly high expression in one cell line, but expression close to the mean in the other cell lines.

To identify genes exhibiting this type of expression pattern, we sorted the z-score expression for the genes in each cell line. The top 200 genes in this sorted list were classified as being up-regulated in that cell type (UR genes). Down-regulated genes exhibit low expression levels in one cell type, but are otherwise constitutively expressed in other cell lines (Figure 5B) (Thorrez et al. 2011). We classified the last 200 genes in the sorted z-score expression list as being the cell-type specific down-regulated genes (DR genes). Constitutively expressed genes (Figure 5C) were identified by filtering all genes that were not in UR and DR gene sets in any cell line and had absolute expression z-score values < 1.7 in all cell lines. Using this cutoff, 168 genes displayed a pattern of constant expression levels across all cell lines.

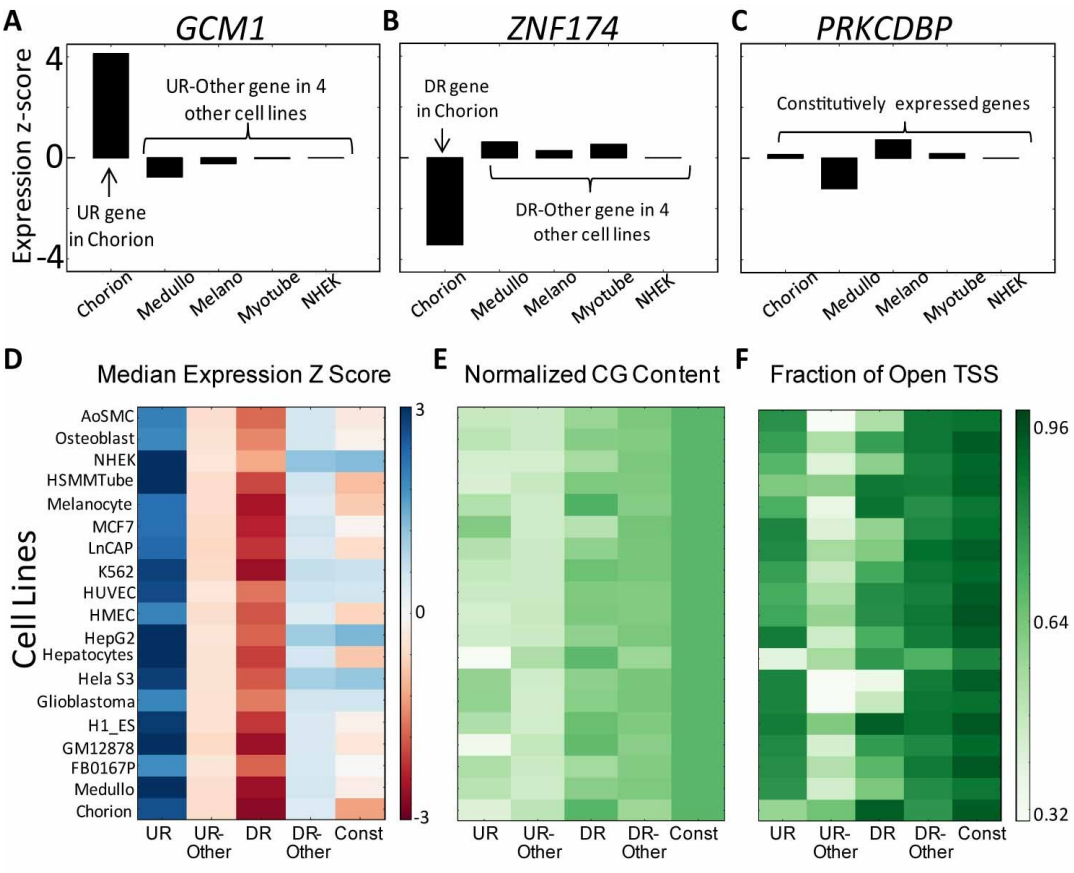


Figure 5: Cell-type specific gene expression and definition of gene classes

(A, B and C) Representative examples of different patterns of gene expression. Note that z-score values are calculated from expression across all 19 cell lines. **(A**) A gene where the expression is specifically up-regulated in the first cell line (UR gene). **(B)** A gene that is specifically down-regulated in the first cell line (DR gene). **(C)** A gene that has low variability in expression (Constitutively expressed gene). **(D)** Median expression z-scores for the genes in each set in each cell line. **(E)** Normalized CG content from the promoter regions of genes. **(F)** The fraction of TSS in each gene set that were in a region of open chromatin. **(E and F)** share the same color map.

To address how up-regulated genes are expressed in one particular cell type, we grouped UR genes from all other cell types and denoted this group as UR-Other genes (Figure 5A). We imposed the additional constraint that such genes would show an expression z-score < 0 in the cell type of consideration, i.e., had expression below its mean expression. As an example, *GCM1* (Figure 5A) was highly expressed in the first cell type and in none of the others shown. It was therefore grouped into the UR class for the first cell type, and into the UR-Other class in each of the other cell types. Similarly, genes denoted as DR-Other had to be classified as down-regulated in another cell line and had an expression z-score > 0 in the cell type of consideration (Figure 5B). In this way, we defined different classes of transcriptionally active (UR, DR-other, and Constitutive) and transcriptionally inactive genes (UR-other, DR) from the point of view of each cell line in comparison to other cell lines.

By definition, UR and DR genes displayed the highest and lowest z-score gene expression, respectively (Figure 5D). UR genes were consequently enriched in functions related to the tissue type of origin (Table 2). DR-other and constitutive genes showed similar expression values and had higher expression values than genes in the UR-other class. To understand whether these different classes had different properties in sequence composition and chromatin state, we first inspected normalized CG content in the proximal promoter region (Figure 5E). UR and UR-Other genes had the lowest average CG content compared to the other classes of genes. Constitutively expressed genes displayed a particularly high CG content in their proximal promoter regions, as

previously reported (Yamashita et al. 2005; Carninci et al. 2006; Zhu et al. 2008); however, this observation clearly extended to the DR and DR-Other gene classes, which had a CG content slightly lower than constitutively expressed genes.

Table 2: GO analysis for UR genes in each cell line.

Cell Line	GO Categories of UR genes
Chorion	Placenta, inflammatory response, extracellular region, cytokine binding
Medulloblastoma	Retina, Brain, visual perception, ion channel complex, gated channel activity
B0167P	Skeletal system development, pattern specification process, extracellular region part, intrinsic to plasma membrane, growth factor activity
GM12878	B-cell, spleen, immune response, cell activation, MHC class II receptor activity
H1_ES	Brain, ion transport, synapse, plasma membrane, gated channel activity
Glioblastoma	Regulation of transcription, DNA-dependent, anterior-posterior pattern formation, DNA binding, zinc ion binding
Hela S3	
Hepatocytes	Liver, acute inflammatory response, complement activation, oxygen binding
HepG2	Liver, lipid homeostasis, cholesterol metabolic process, sterol homeostasis, extracellular space
HMEC	Keratinocyte, ectoderm development, epidermis development, epithelial cell differentiation
HUVEC	Umbilical Vein Endothelial cell, angiogenesis, vasculature development, plasma membrane part, cell adhesion
K562	Blood, platelet, hemopoiesis, intrinsic to plasma membrane
LnCAP	Prostate, Prostatic carcinoma, synaptic transmission
MCF7	
Melanocyte	Skin, Melanoma, pigmentation during development, melanocyte differentiation, melanosome
HSMMTube	Skeletal muscle, heart, muscle system process, muscle tissue development, structural constituent of muscle
NHEK	Keratinocyte, keratinocyte differentiation, epithelial cell differentiation, desmosome
Osteoblast	Fibroblast, Osteoblast, skeletal system development, extracellular structure organization
AoSMC	Fibroblast, response to wounding, cell adhesion, extracellular region part, chemokine activity

The UP_Tissue entries from DAVID were used to identify the similarity of expression to known tissue types.

Constitutively expressed and DR-Other genes had the highest proportion of their TSS in regions of accessible chromatin (Figure 5F). DR genes displayed slightly lower chromatin accessibility compared to DR-Other, indicating that repression of DR genes largely occurs while maintaining chromatin accessibility at the TSS. UR genes also showed a high proportion of genes containing an accessible TSS, at similar levels to DR and DR-Other. In stark contrast, UR-Other genes had the lowest fraction of TSSs that overlapped a DHS. These results indicate that even though UR-other and DR genes are both transcriptionally inactive in a cell type of consideration compared to other cell types, they are likely to be regulated via different chromatin remodeling mechanisms. Specifically, genes that are up-regulated in a small number of cell types likely maintain a closed chromatin conformation until cellular processes require up-regulation. In contrast, down-regulated genes may be viewed as constitutively expressed genes that are repressed in a single cell type. UR genes had intergenic and gene body DHS associated with them (Figure 6), in agreement with previous results indicating that cell-type specific expression is mediated by distal *cis*-regulatory regions (Song et al. 2011). Overall, these results indicate that different classes of transcriptionally active and inactive genes have different CG content and chromatin accessibility at their TSS.

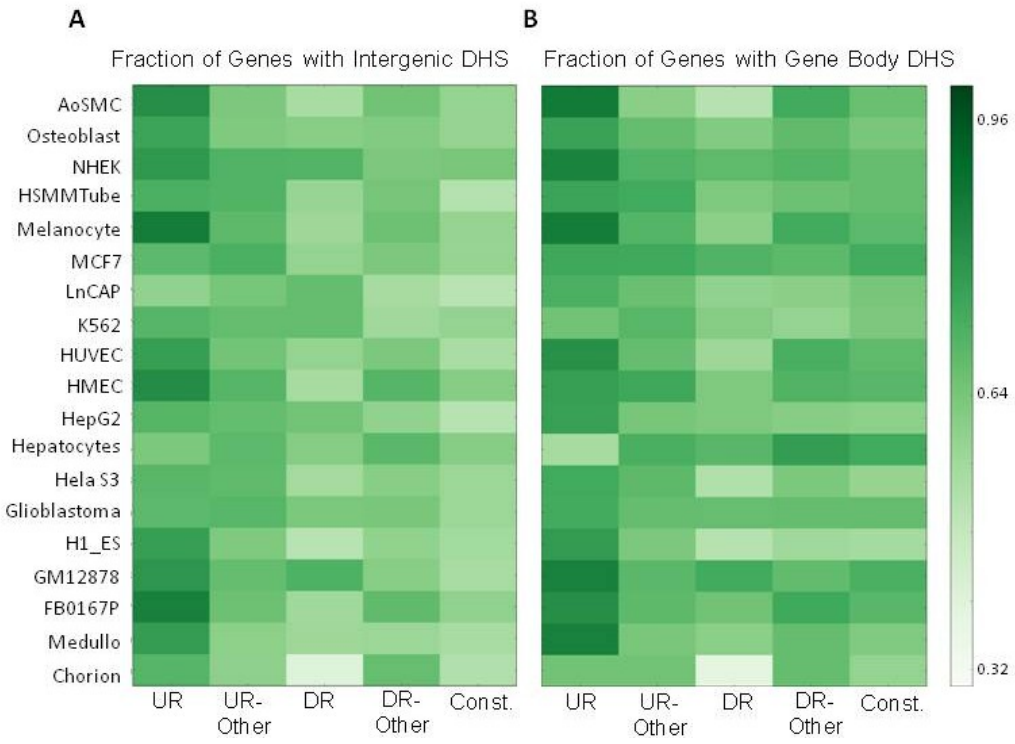
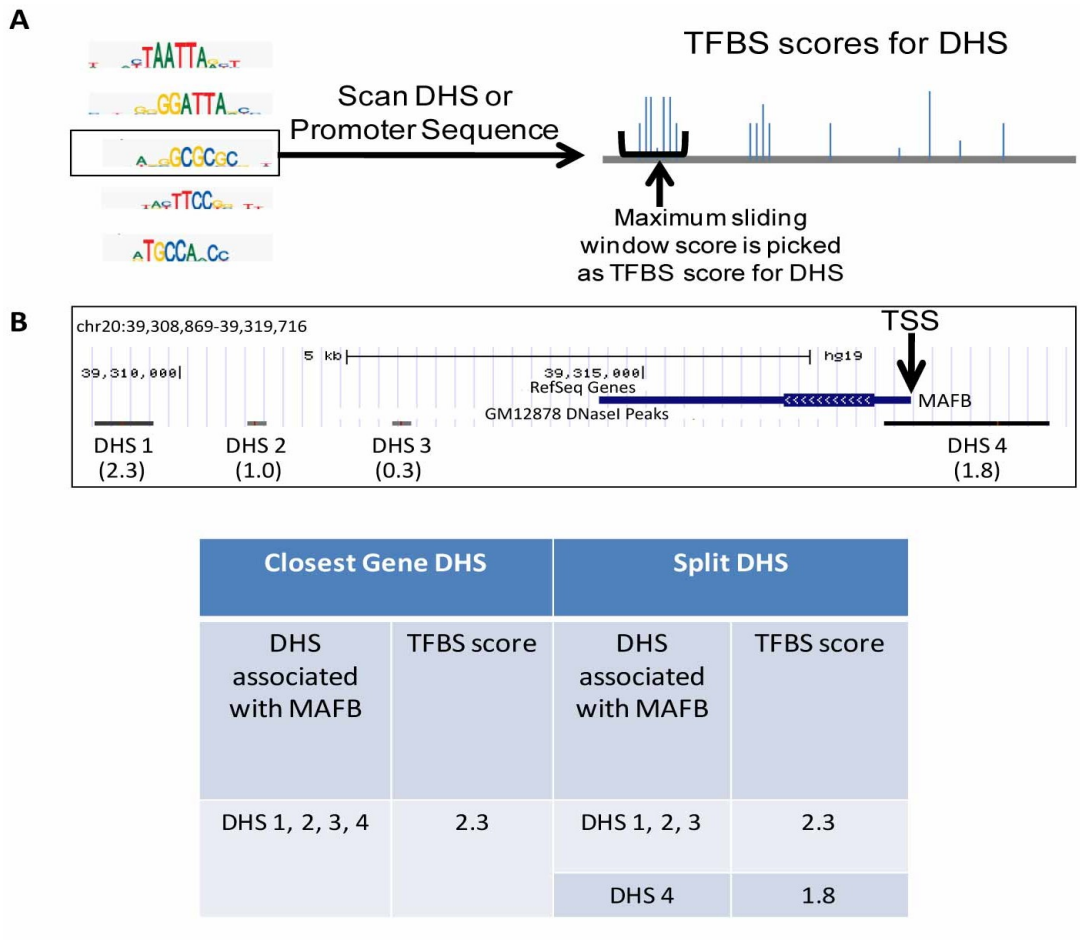


Figure 6: The fraction of genes in each gene set that had Intergenic DHSs and Gene Body DHS

3.3.4 Classifying tissue-specific expression from sequence features in open chromatin

To predict gene expression patterns from sequence, approaches have frequently used features contained within fixed-size proximal promoter sequences. We used DHS data from a large number of cell types to determine whether using both proximal and distal regulatory regions with open chromatin would improve predictive models for cell-type specific expression patterns. Position Weight Matrices (PWMs) for TFs in vertebrates were compiled from Transfac, JASPAR and UniProbe databases (Matys et al. 2006; Bryne et al. 2008; Newburger and Bulyk 2009). For each DHS, 789 PWMs were used to calculate TFBS scores that accounted for local dinucleotide composition. The maximum sliding window score for each PWM was used as the TFBS score for that DHS (Figure 7A). To associate DHS with specific genes that they are likely to regulate, we applied a simple approach of associating each DHS with the closest TSS (closest gene DHS). For each TF, we then chose the maximum TFBS score across all DHS associated with a gene (Figure 7B). As an alternative approach, we split DHS into distal sites (a set including both Gene-Body and Intergenic DHS) and TSS DHS sites and used the maximum TFBS in each set as individual features (split DHS). This doubled the number of features and allowed us to identify different characteristics of TSS-overlapping vs. distal DHS. To compare our models to previous approaches, we also used TFBS features calculated in proximal promoters, defined here as -900 to +100 nucleotides surrounding the TSS (Landolin et al. 2010).



Closest Gene DHS		Split DHS	
DHS associated with MAFB	TFBS score	DHS associated with MAFB	TFBS score
DHS 1, 2, 3, 4	2.3	DHS 1, 2, 3	2.3
		DHS 4	1.8

Figure 7: Transcription factor binding site as features

(A) DHS and promoter sequences are scanned with PWMs. TFBS scores are log-likelihood ratios of PWM over the background model. A sliding window is used to identify the score for each DHS or Promoter. **(B)** Example to show association of DHS with genes. Numbers in the brackets are example TFBS scores for the DHS for a specific DHS. Two methods of association were used. In closest gene DHS, DHS 1- 4 from the GM12878 cell line are associated with the gene MAFB. For the TF in consideration, the maximum of all TFBS scores is 2.3. In Split DHS, we separated DHS overlapping the TSS and other DHS. This resulted in two features for each gene for each TF.

We used the TFBS scores as features for sparse logistic regression classifiers to discriminate between different gene classes. These classifiers balance the use of many available features against model complexity, effectively selecting a small subset of informative features which are used in the classification. We trained cell-type specific classifiers on the task to discern whether a gene belonged to a specific expression pattern (e.g., UR vs. UR-Other, UR vs. DR, UR vs. constitutive, etc.). The area under the Receiver Operating Characteristic curve (AuROC) metric was used to evaluate the performance of a model, where a value of 0.5 indicates random assignments and 1.0 indicates perfect classification. To not bias results due to different amounts of training data, the positive sets of up- and down-regulated genes were all of the same size.

The performance of the classifier using only proximal promoter information is close to that of a random classifier, across all tasks. All the classifiers using DHS sequences display strong improvements in performance over this baseline in discriminating genes that are up-regulated in different cell types (UR vs. UR-Other, Figure 8A), with a greater improvement in performance coming from the Split DHS approach with separate features for the TSS and Distal DHSs (median AuROC ~0.73). Similar results were obtained when training classifiers to distinguish between specifically up- and down-regulated genes from the same cell types (UR vs. DR, Figure 8B), and to distinguish up-regulated from constitutively expressed genes (UR vs. Const., Figure 8C). Discriminating down-regulated genes from different cell types (DR vs. DR-Other), and down-regulated from constitutively expressed genes (DR vs. Const.),

resulted in lower accuracies but still showed the trend of better performance with DHS compared to proximal promoter sequence (Figure 9). All results clearly indicate that strong performance improvement is achieved by scanning for TFBS matches in open chromatin regions.

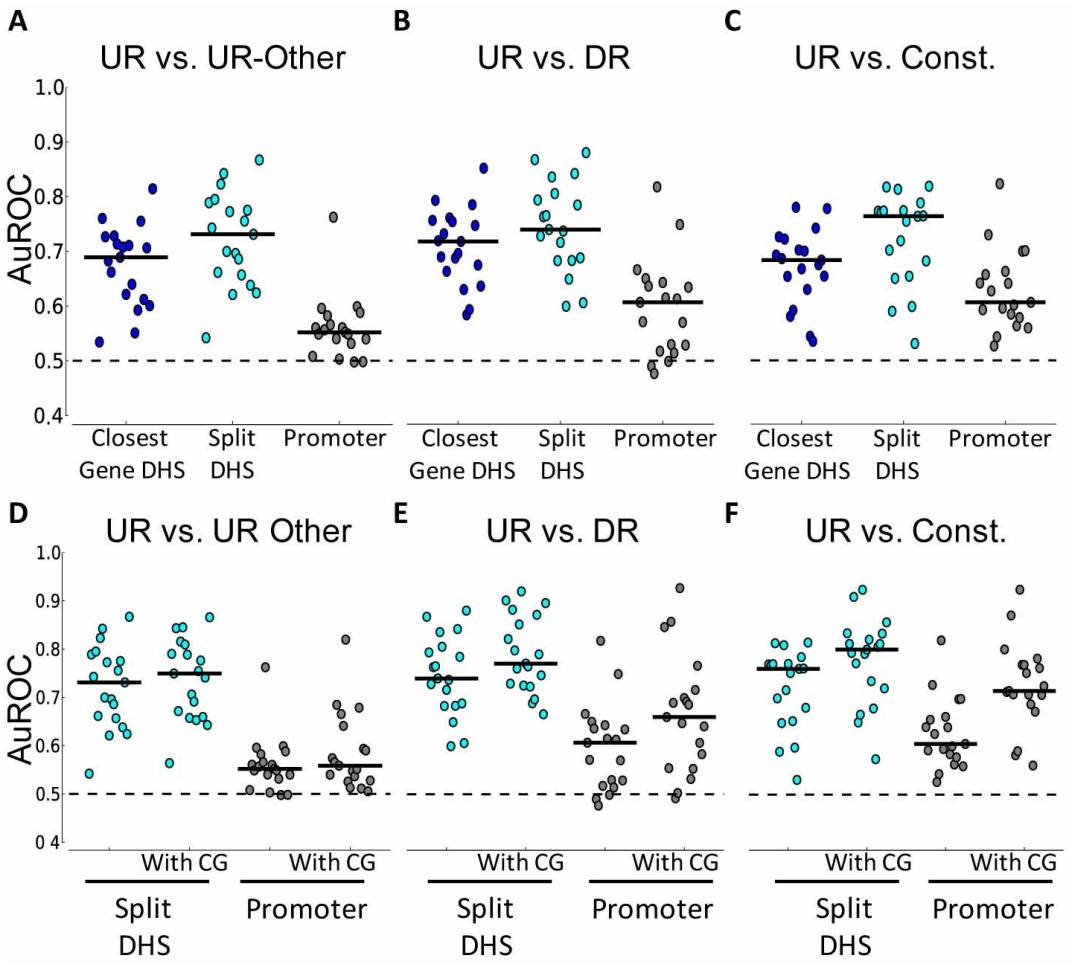


Figure 8: Classifier performance for various classification tasks with UR genes

(A, B, C) Performance of the classifier using all PWMs. Each figure compares the performance of 2 methods of associating DHSs to genes (Closest Gene DHS and Split DHS) with the proximal promoter. The solid black lines across the dots indicate the median. Across all figures, the promoter sequence classifier does not perform as well as the performance achieved by using closest gene DHS and Split DHS and is significant at the 0.05 level (paired t-test). **(D, E, F)** Impact of normalized CG dinucleotide content on classifier performance. Results using the Split DHS and Promoter sequence are shown. Without CG data is same as in A-C. All figures show average results from 5 iterations of

4-fold cross-validation. The dotted line indicates an AuROC of 0.5 which is the performance of a random classifier.

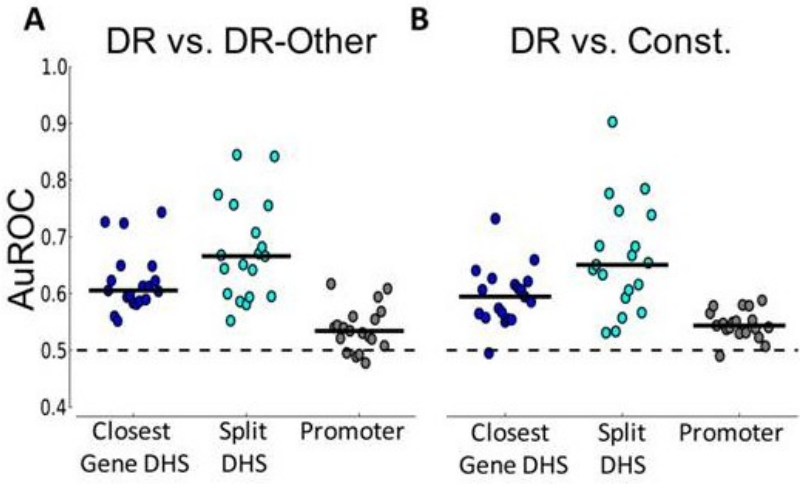


Figure 9: Classifier performance for various classification tasks with DR genes

(A, B) Performance of the classifier using all PWMs. Each figure compares the performance of 2 methods of associating DHSs to genes (Closest Gene DHS and Split DHS) with the proximal promoter. Across all figures, the promoter sequence classifier does not work as well as the performance achieved by using closest gene DHS and Split DHS and is significant at the 0.05 level (t-test). The black lines across the dots indicate the median. All figures show results from 5 iterations of 4-fold cross-validation. The dotted line indicates an AuROC of 0.5 which is the performance of a random classifier.

3.3.5 Evaluating the influence of CG dinucleotide content

CG dinucleotide content in the proximal promoter sequence of genes is a common sequence feature that is directly or implicitly used to distinguish various classes of genes. Adding CG dinucleotide content as an additional feature led to a variable impact on classifier performance depending on the classification being considered (Figure 8D-F). Specifically, when using open chromatin information, adding CG content did not substantially improve the performance of classifying UR genes from UR-Other genes (Figure 8D). In the case of the Split DHS, only 6 of the cell lines had a significant coefficient for the CG dinucleotide coefficient (mean across all cell lines = -0.66, std. dev. = 1.40; coefficient was set to 0 when not significant). Due to the means being close to zero and the standard deviations being large, the effect of using CG content to discriminate between UR and UR-Other genes was largely negligible. Only for few cell types, such as hepatocytes, we observed a negative regression coefficients of significant magnitude (-5.11 for Split DHS and -3.31 for Promoters). This is in agreement with previous results showing that liver specific genes have promoters with lower CpG content (Smith et al. 2005).

As anticipated by the trends observed in Figure 5E, UR vs. DR classification tasks benefitted more by the addition of CpG content. Here, this feature was deemed to be significant in 17 of the cell lines for the Split DHS (Figure 8E). Further, the regression coefficients were largely negative which indicated the higher CG content among the DR genes (mean = -2.88, std. dev = 1.55). As has been shown before (Yamashita et al. 2005;

Carninci et al. 2006; Zhu et al. 2008), we observed that high CpG content is predictive of constitutively expressed genes when compared to UR genes (Figure 8F). The regression coefficient for the feature was significant in all cases (mean = -3.25, std. dev 1.44). The CpG content feature had almost no impact in classifying DR from DR-Other genes (Figure 9). Finally, CpG content had a significant coefficient in classifying DR from Constitutively expressed genes in only 1 cell line (mean = -0.09, std. dev. = 0.39).

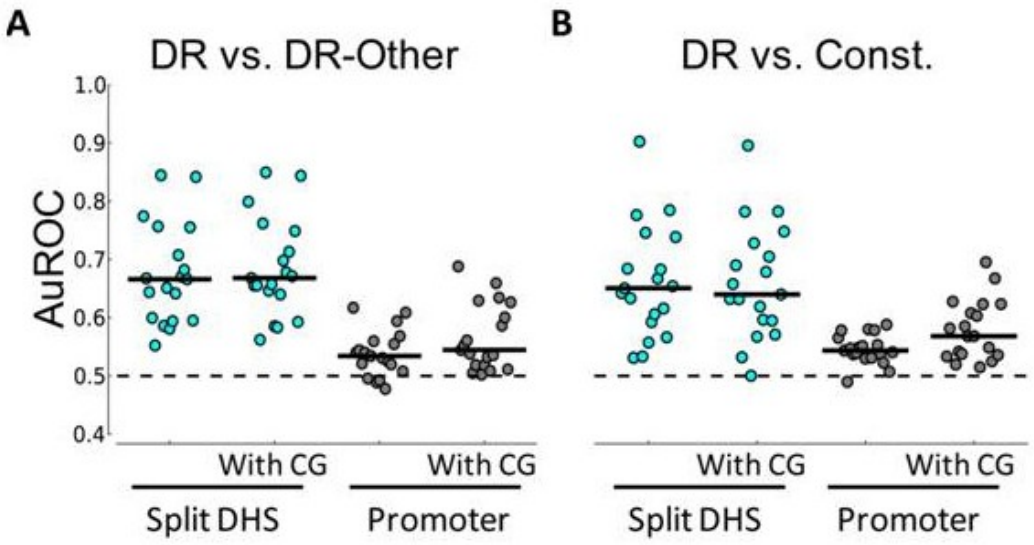


Figure 10: Impact of normalized CG dinucleotide content on classifier performance for discriminating DR genes.

(A, B) Results using the Split DHS and Promoter sequence are shown. The dotted line indicates an AuROC of 0.5 which is the performance of a random classifier. CG content is shown to help for the proximal promoter where other factors are not informative.

Adding CG content to the baseline proximal promoter models reconciled the apparent discrepancies between previous studies and results reported in Figure 8A-C, as all classification tasks were improved upon for the proximal promoter. However, the DHS models with CG content outperformed baseline proximal promoter models with the inclusion of CG content (paired t-test < 0.05). In fact in all cases except UR vs. constitutive genes, DHS models even without CG content perform significantly better than both proximal promoter models (paired t-test < 0.05). Note that while adding CG content provided enormous performance gains for certain classification tasks (UR genes vs. Constitutive genes) this could be considered misleading. If TFBS scores are not explicitly normalized for local nucleotide composition, as we have done here, decent performance results can be achieved based solely on the different CG content observed for down-regulated and constitutively expressed genes compared to up-regulated genes. CG content is predictive in the case of classifying constitutive and DR genes from UR genes, but is not very useful in differentiating between genes that are up- or down-regulated in different cell types. It is notable that the categories that are less aided by CG content are exactly those where our classifiers displayed the most predictive value.

3.3.6 Identifying candidate regulators

In addition to classifying genes belonging to different groups, we inspected the classifiers to identify motifs that were most informative in the classification task, i.e., those PWMs that had large regression coefficients (Table 3). This identified several TFs

with known impact on transcriptional output in the cell line of interest. For example, *YY1*, *SPI1* and *IRF8* are crucial in the specification of B-cells (GM12878 cell line) (Lu et al. 2003; Liu et al. 2007; Sokalski et al. 2011). We also identified the *REST* motif as a positive regulator of UR genes in medulloblastoma cell line that is of neural origin (Table 3). *REST* specifically down-regulates neuron-specific genes in many non-neuronal cell lines, and its expression is suppressed in neurons (Schoenherr and Anderson 1995). As a result, the model identified the *cis*-elements that are present in the DHS associated with neuron specific genes as the factor that separates these genes from the genes up-regulated elsewhere. This example illustrates that the inactivation of a repressor can also explain up-regulation of genes. Other well characterized factors included *ETS1* in HUVEC cells and *HNF4A* for HepG2 cells (Cereghini 1996; Oda et al. 1999; Yordy et al. 2005).

Table 3: Results for each cell line using Split DHS from All TFs for the UR – UR Other and UR – DR classification task.

Cell Type	UR – UR Other Genes			UR – DR genes		
	AuROC	TFs - Positive Coefficient	TFs - Negative Coefficient	AuROC	Positive Coefficient	Negative Coefficient
Chorion	0.54		*ZNF143*	0.87	*HBP1*	*E2F1, NFYA, GABPA*
Medulloblastoma	0.79	*PDX1, CRX, REST*		0.80	*MEF2A, CRX, REST*	
FB0167P	0.74	*DMRT1, JDP2*		0.73	*POU2F1*	
GM12878	0.79	*YY1, NFE2L1-MAFG, SPI1, IRF8*	*E2F3, E2F4-TFDP2, E2F1*	0.76	*INSM1, IRF8*	*AHR-ARNT, HINFP*
H1_ES	0.66			0.77	*NANOG*	*TFAP2A, GABPA, ELK4*
Glioblastoma	0.82	*ZNF143*		0.74	*STAT5B*	
Hela S3	0.84	*USF1, GABPA*		0.84		
Hepatocytes	0.70		*RFX1, ZFP161, FOXN1*	0.80	*NR2F2, RXRA-NR1H2*	*E2F1, FOXN1*
HepG2	0.77	*ZEB1*		0.68	*HNF4A*	
HMEC	0.62			0.72		*ELK4*
HUVEC	0.70	*ETS1, SPI1*		0.74		*ELK4, E2F*
K562	0.69	*YY1, LMO2 bound to TAL1, TCF3 and GATA1, ETS1*		0.60		
LnCAP	0.66			0.65	*SOX5*	
MCF7	0.76	*GATA6, ZFX, TCF3, HINFP*		0.68	*RBPJ*	
Melanocyte	0.78	*SREBF1, MEF2A, AHR-ARNT*		0.84	*MYCN, ELK4*	*GABPA*
HSMMtube	0.64	*MEF2A, PKNOX2*	*ZNF423, PAX6, PAX3*	0.79	*BACH2*	*NFYA, GABPA*
NHEK	0.73	*ZNF410*		0.69	*MAF, MTF1*	*FOXD1*
Osteoblast	0.62			0.61	*MYB*	*FOXN1*
AoSMC	0.87	*DMRT1, CEBPB, PPARG*		0.88	*CEBPB, PATZ1*	*NFYA*

TFs with positive and negative coefficients are shown for both sets of TFs used.

The feature set described thus far was comprehensive in that it used available PWM information from multiple sources, independent of the expression levels of transcription factors or the potential redundancy of features. To assess how much cell-type specific regulation can be explained by the cell-type specific expression of transcription factors themselves, we selected the top 10 TFs with highest absolute z-scores from each cell line and had PWMs that were not similar to each other (Table 4).

Table 4: Top 10 non-redundant TFs with highest absolute expression z-score in each cell line.

Cell Line	Cell-Type Specifically Up-Regulated TFs	Cell-Type Specifically Down-Regulated TFs
Chorion	*ASCL2, EGR1, OSR2, GCM1, DLX5*	*IRF3, E2F3, ZFP161, USF1, ELK1*
Medulloblastoma	*CRX, INSM1, NHLH1, HLX, SIX3, SOX11*	*REST, SMAD3, NR2F2, SP100*
FB0167P	*ZBTB12, STAT1, LHX9, ZIC1, OSR1, HOXC11*	*AIRE, RFX3, IRF5,BACH2*
GM12878	*HIC1, ARID5A, IRF4, SPIB, EGR2, POU2F2*	*FOXP1, GLIS2, TCF7L2, BCL6*
H1_ES	*ZBTB3, MYCN, SOX21, SOX11, ZIC3, SOX2, OTX2*	*MEIS1, NR2F2, HOXC9*
Glioblastoma	*IRF3, HOXD10, HOXB5, NKX3-2, PAX6, ZIC1, PITX2, HOXD11*	*AHR, STAT5A*
Hela S3	*ESRRA,E2F2, PAX6, ARNT, SP1, MAFK, ELK1, FOXF2, ATF1, MEOX1*	
Hepatocytes	*BACH2, HNF4A, RXRA, AR, STAT3*	*FOXJ3, E2F3, RFX7, SIX4, HOXA6*
HepG2	*CEBPA, HNF1A, GFI1, HNF4A, HOXD1, NFYA, FOXA2, HOXA3, TCF7, SOX9*	
HMEC	*STAT4, IRF6, EGR3, OSR1, HOXA5*	*ZBTB3, DR1, HOXA11, SIX1, HOXA5, PAX2*
HUVEC	*SOX18, GBX2, HIC1, ARID5A, SOX17, HOXA3, BCL6B, HOXA9*	*ZBTB6, BACH1*
K562	*WT1, HOXB9, GFI1B, ESRRB, STAT5A, LEF1, MYB, GATA1*	*ARX, KLF7*
LnCAP	*ZBTB7B, HOXC6, AR, NKX3-1, MAFB, ELF5, HOXB13*	*FOXJ1, STAT6, KLF7*
MCF7	*ESR1, SPDEF, LMX1B, IRX5, GSC, MSX2, GATA3*	*SOX14, POU6F1, FOXI1*
Melanocyte	*MAF, IRF4, PAX3, TBX5, IRX6, LEF1*	*NKX3-1, BBX, GABPA, CUX1*
HSMMTube	*MYF6, SOX11, SIX1, PITX2*	*ZBTB3, ZBTB12, HOXD13, GATA6, STRA13, HLXB9*
NHEK	*FOXJ2, VDR, MTF1, SOX15, EHF, SOX8, HOXA1, MAF, IRX4, GATA5*	
Osteoblast	*STAT4, STAT1, EGR2, BACH1, SIX1, HOXA11, GLIS2, BARX1, PROP1*	*SOX21*
AoSMC	*OSR1, MEIS1, HBP1, CUX1,*	*POU3F2, HOXC11, PAX4, SOX8, PITX3, SOX7*

Using sparse logistic regression classifiers trained on these small sets of variables, we observed similar predictive trends, which indicated that a subset of cell-type specific TFs were predictive of tissue-specific expression (Figure 11). Using only promoter CG content or the status of the chromatin at the TSS as features for a baseline comparison shows that motifs in DHS regions significantly contribute to the performance improvement across all comparisons. In addition, we used genomic regions identified as conserved in the 46-way placental mammal phastCons track from the UCSC genome browser. We note that using conserved sequences and particularly conserved non-coding regions improved performance compared to the promoter. However, the AuROC was still highest when DHS sequences were used indicating the presence of motifs in weakly conserved DHS regions contribute to the performance improvement. Finally, to assess the potential influence of insulators, we excluded DHS that overlapped *CTCF* binding sites for classifiers trained specifically for the 9 cell types for which genome-wide *CTCF* ChIP data was available (Figure 12). While this did not impact classification of UR genes, it reduced the accuracy of identifying DR genes, demonstrating that regions containing insulator sites are likely to contain regulatory information for the repression of genes.

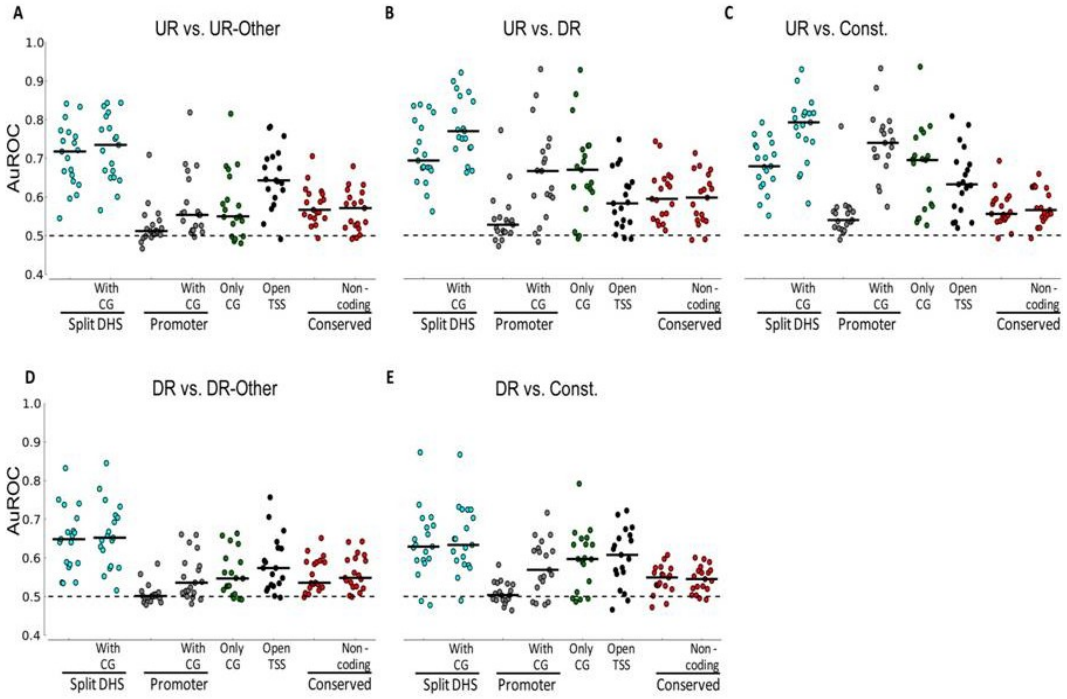


Figure 11: Performance of classifier using top 10 non-redundant highest absolute z-score

Each figure compares the performance of Split DHS with the proximal promoter. Across all figures, the promoter sequence classifier does not work as well as the performance achieved by using Split DHS. The black line across the dots indicates the median. In addition, the impact of CG content, and knowing the whether the promoter is open is shown. All figures show results from 5 iterations of 4-fold cross-validation. The dotted line indicates an AuROC of 0.5 which is the performance of a random classifier.

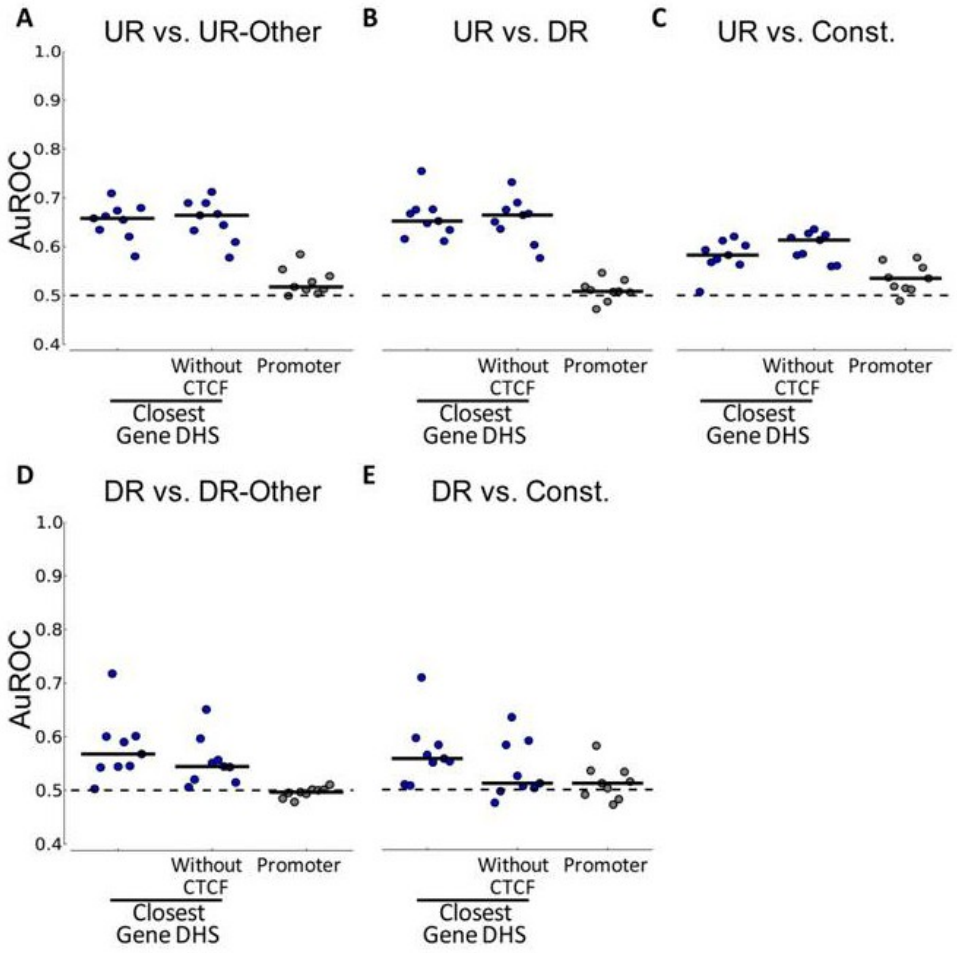


Figure 12: Performance of classifier using top 10 non-redundant highest absolute z-score using DHS with or without CTCF

Each figure compares the performance of 2 methods of associating DHSs to genes using Closest Gene DHS and the inclusion and exclusion of DHS with CTCF ChIP-seq peaks and with the proximal promoter. Note that only 9 cell lines had CTCF ChIP-seq data. All figures show results from 5 iterations of 4-fold cross-validation. The dotted line indicates an AuROC of 0.5 which is the performance of a random classifier.

Knowing both the regression coefficient in our model and the expression level of a potential regulator provided clues as to whether the TF in question is an activator or a repressor in the cell line, as highlighted for *REST* in medulloblastoma cells (Table 5). As another example, *NR2F2* was identified as a positive predictor of up-regulated genes for embryonic stem cells. However, *NR2F2* is a known negative regulator of *POU5F1*, a critical gene involved in pluripotency (Rosa and Brivanlou 2011). As expected, *NR2F2* is down-regulated in ES cells. We also identified other known positive regulators, such as *GATA1* in K562 cells, (Huang et al. 2005) and *MYF6* in Myotubes (Fan et al. 2011). Note that genes that have both positive and negative coefficients have different effects when in TSS DHS and Distal DHS.

Table 5: Classifier performance for each cell line using Split DHS from Top 10 highest absolute z-score of expression and non-redundant TFs

Cell Type	UR – UR Other Genes			UR – DR genes		
	AuROC	TFs - Positive Coefficient	TFs - Negative Coefficient	AuROC	TFs - Positive Coefficient	TFs - Negative Coefficient
Chorion	0.55		***OSR2***, *ELK1*	0.83	*ZFP161*	*E2F3, ELK1*
Medulloblastoma	0.77	***CRX***, *REST*		0.8	***CRX***, *REST, NR2F2,* ***SOX11***	
FB0167P	0.72	*BACH2,* ***ZIC1***, *AIRE*		0.72	***STAT1, ZBTB12,*** *BACH2,* ***HOXC11, ZIC1,*** *AIRE*	
GM12878	0.75	***EGR2, SPIB***		0.64	***SPIB***	***ARID5A***
H1_ES	0.67	***ZIC3, OTX2***, *NR2F2*		0.69	*NR2F2,* ***ZIC3, OTX2***	*MEIS1, NR2F2*
Glioblastoma	0.81	***ZIC1, IRF3, BAPX1***		0.74	***ZIC1, HOXD10***	
HelaS3	0.84	***PAX6, E2F2, FOXF2, ELK1, ARNT, ESRRA***		0.84	***MEOX1, ARNT, FOXF2, PAX6, ELK1, ESRRA***	
Hepatocytes	0.70	*FOXJ3* **, HNF4A, RXRA, STAT3**	***RXRA***, *RFX7, HOXA6, FOXJ3, E2F3,* ***STAT3***	0.78		*E2F3*
HepG2	0.77	***GFI1, HNF4A***		0.67	***SOX9, FOXA2, HNF4A***	
HMEC	0.6	***STAT4***		0.68	***STAT4, IRF6***	
HUVEC	0.67	***SOX17***		0.71	***HIC1, SOX17***	
K562	0.66	***GATA1***		0.64	***GATA1***	
LnCAP	0.64	***NKX3-1***		0.6	***ZBTB7B***	
MCF7	0.74	***GATA3***		0.68	***GATA3, ESR1***	
Melanocyte	0.76	***MAF, LEF1, IRF4,*** *CUX1, GABPA*		0.83	***IRF4***, *GABPA*	***TBX5***, *GABPA*
HSMMtube	0.61	*HLXB9,* ***MYF6***, *ZBTB3*	***SOX11***, *SIX1, ZBTB12, STRA13, ZBTB3*	0.67	***MYF6***	***MYF6***, *STRA13, ZBTB12,* ***SOX11***, *GATA6, ZBTB6*
NHEK	0.72	***MTF1, MAF***		0.67	***MTF1***	
Osteoblast	0.63	***BACH1, STAT4, GLIS2***		0.56	***STAT4, BACH1***	
AoSMC	0.83	***MEIS1, OSR1,*** *HOXC11, SOX8, PITX3, PAX4*		0.82	*PITX3,* ***MEIS1, OSR1,*** *HOXC11*	

UR vs. UR-Other and UR vs. DR classification tasks are shown. TFs with positive and negative coefficients are shown for both tasks. Genes in bold are up-regulated and other genes are down-regulated in the cell line. Several of the same factors help in classifying UR vs. UR-Other genes and UR vs. DR genes.

For *HNF4A* in HepG2 and *GATA1* in K562 cells, ChIP data is available from the ENCODE project. To validate the predictions made by our model, we looked for overlap of these ChIP sites with DHS sites associated with different sets of genes. In HepG2 cells, 19% of all genes with an associated DHS overlapped a *HNF4A* binding site. Strikingly, 64.5% of the UR genes had a DHS overlapping an *HNF4A* ChIP peak (p-value<1e-12, binomial test). Conversely, only 10.5% of DR genes had a DHS that overlapped an *HNF4A* site (p-value<1e-3). In K562 cells, we found that 6% of all genes had an associated DHS with a *GATA1* ChIP peak. However, 31.5% (p-value<1e-12) of UR genes and only 3.5% (p-value<0.1) of DR genes had a DHS with a GATA ChIP peak. The ChIP binding data provided strong and independent evidence that our models identify relevant factors that regulate the transcriptional program in these cells.

In addition, we investigated the accuracy of our predictions of TFBS locations in DHS. In HepG2, 5215 of the 6597 *HNF4A* ChIP peaks overlapped the predicted TFBS in DHSs. Furthermore, using TFBS scores led to a high accuracy on discriminating between positive and negative sets defined by ChIP peaks (AuROC of 0.79). In K562 cells, only 315 of the 1704 *GATA1* ChIP peaks overlapped the predicted TFBS in the DHS; yet, the AuROC still remained high at a value of 0.88. This indicated that high scoring TFBSs accurately predicted binding of *GATA1* to these sites. We note that the low percent overlap may arise from non-specific or indirect binding of *GATA1*.

To assess the presence of additional sequence motifs not accounted for by the sets of known PWMs, we used the discriminative version of MEME to perform motif

finding (Bailey et al. 2010), identifying motifs differentially enriched between UR and UR-other respectively DR genes (Table 6). While some of the identified motifs corroborated the importance of features from the set of top 10 TFs (*FOXA2* [formerly *HNF3B*] in HepG2), others corresponded to TFs that were not in this list. These are candidate TFs that are not among the most differentially expressed, but still might be involved in the transcriptional program, potentially through other steps of activation. We note that we largely did not recover the motifs recently identified in a subset of 7 of the 19 cell lines (Song et al. 2011). In contrast to this study, which used the sequences from cell-type specific DHS as foreground and the subsets of cell-type specific DHS in other cell types as background, we analyzed the sequences from all DHS associated to a gene, and defined the background according to the classification tasks .

Table 6: Matches to motifs identified using MEME.

Cell Line	UR – UR Other Genes	UR – DR Genes
Chorion	*SPI1*	*FOXP1, SP1*
Medulloblastoma	*TAL1-GATA1*	*EWSR1-FLI1*
FB0167P	*SPI1*	*SPI1*
GM12878	***SPIB*** *, ***FOXP1***	***SPIB*** *, ***FOXP1***
H1_ES	***ZIC3*** *,*EWSR1-FLI1, FOXP1*	*EWSR1-FLI1,SP1, FOXP1*
Glioblastoma	*EWSR1-FLI1, NFE2L2*	*IRF*
Hela S3	*EWSR1-FLI1, TAL1-GATA1,* ***FOXF2*** *	***SP1, FOXF2*** *
Hepatocytes	***STAT3*** *	***FOXJ3*** *
HepG2	*SP1, ZNF219,* ***FOXA2***	*SP1,* ***FOXA2*** *
HMEC	*SPI1*	*ETS2*
HUVEC	*ETS2, FOXP1, ZBTB7B*	*EWSR1-FLI1, ZBTB7B*
K562	*FOXP1, SP1,* ***WT1***	***WT1,*** *FOXP1*
LnCAP	***ELF5***	***ELF5***
MCF7	***FOXI1***	***GATA3*** *
Melanocyte	*IRF, SP1*	*IRF,* ***IRX6***
HSMMTube	*IRF*	*IRF*
NHEK	*ZNF219,* ***FOXJ2****, SP1*	*SP1,****FOXJ2***
Osteoblast	*IRF, SP1*	*IRF*
AoSMC	*EWSR1-FLI1, ZNF281, NFE2L2, FOXP1*	***PAX4, SOX7****, IRF*

Motifs were first compared to the top10 non-redundant TFs using STAMP. The matches found to that list are shown in bold. * indicates that the TF also was identified in the classifier as being predictive.

3.3.7 Footprints in DNase-seq data show evidence of direct TF binding

While we have shown that the presence of sequence motifs in DHS regions is predictive of cell-type specific gene expression patterns, we were only able to validate the direct binding of two factors due to the lack of ChIP data. This issue is likely to arise in several studies where ChIP data is not available to provide evidence of direct binding of a TF to its cognate binding site. However, if a region is bound by a TF, the profile of aggregate reads around the TF binding site will show that region to be protected from digestion by DNaseI, resulting in a DNase "footprint". Based on this pattern, DNase-seq data has been recently used to identify the precise binding locations of several TFs at base pair accuracy (Hesselberth et al. 2009; Boyle et al. 2011; Pique-Regi et al. 2011).

To assess whether the factors that had high regression coefficients in the classification tasks showed such distinctive footprints, we compiled the DNase-seq reads in a 100bp window centered on the top motif matches across the genome. We expected to see a distinct pattern in the cell line in which a motif was predictive of gene expression. As a control, DNase profiles were compiled for cell lines in which the model did not have a high regression coefficient for the TF. The motif matches used here were chosen to reflect the genome-wide binding of the factor, as opposed to the specific binding sites used to model gene expression.

For several factors, we observed indicative footprints in the region of the motif (Figure 13). For example, *CRX* was predictive of UR genes in the medulloblastoma cell

line, and it exhibited a protected region at the motif (Figure 13A). Importantly, in other cell lines such as GM12878, LnCAP and MCF7, the *CRX* motif did not display a similar level of protection. While *CRX* has been shown to be expressed in certain types of medulloblastoma sub-types (Kool et al. 2008), other factors such as *OTX2* have nearly identical PWMs and are known to be important for transcriptional regulation in medulloblastomas (Bunt et al. 2011). This highlights a caveat in predicting expression from motifs; while we can identify biologically relevant motifs, this type of analysis only suggests a subset of factors that likely bind to a specific motif.

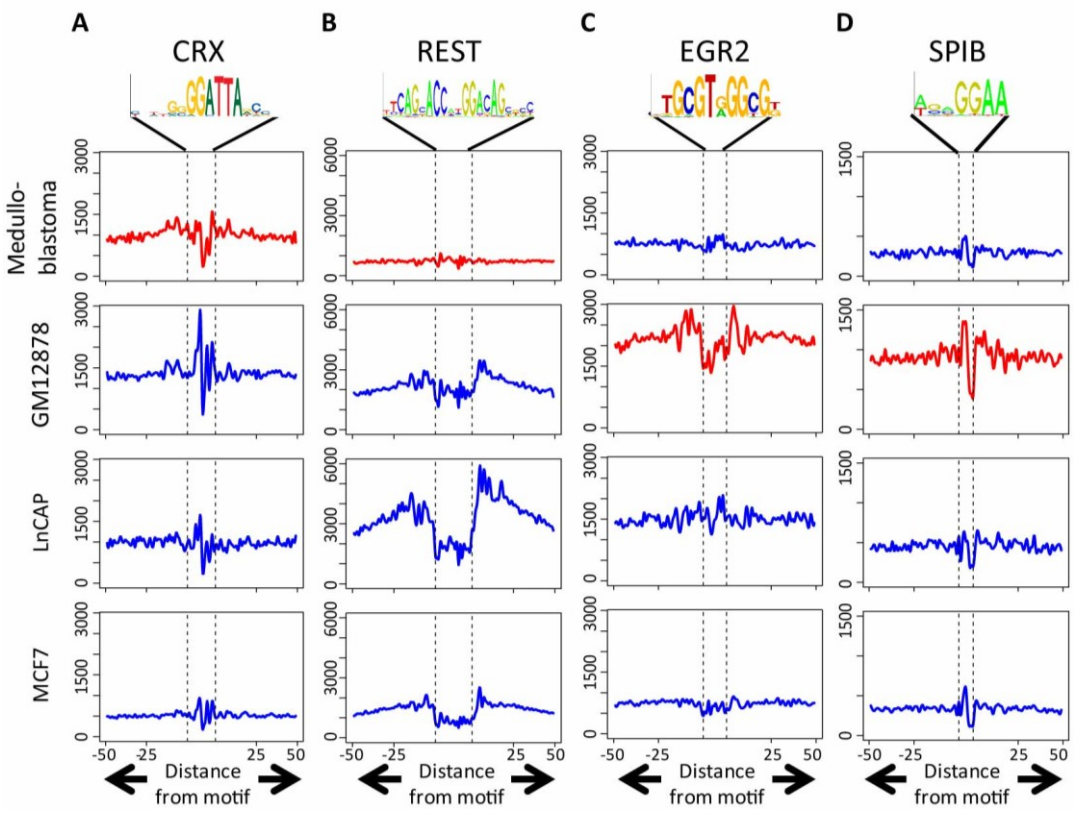


Figure 13: Aggregate plots of DNase-reads around motifs for factors with high regression coefficients.

(A) CRX shows a footprint in medulloblastoma but not in the other cell lines shown. **(B)** REST shows a footprint in other cell lines but not in medulloblastoma where it is not expressed. **(C, D)** EGR2 and SPIB show footprints in GM12878 cell line. Red lines indicate the cell line in which TF is identified as a regulator.

As mentioned earlier, we identified *REST* as a regulator in the medulloblastoma cell line. Since it is not expressed in this cell line, we observed the absence of a footprint in that cell line, and a visible footprint in other cell lines (Figure 13B). Additional footprinting evidence is detected for *EGR2* and *SPIB* in the GM12878 cell line (Figure 13C-D); however, the *SPIB* motif also exhibits a smaller footprint in other cell lines. This could be due to expression of other factors that bind to a similar motif in this cell line. Further work is needed to rigorously quantify these encouraging observations.

3.4 Discussion

In this study, we proposed a new method to predict gene expression by using DNase-seq data from 19 human cell lines. Unlike other strategies that require multiple ChIP-seq datasets for highly informative regulatory factors, a single DNase-seq experiment identifies most regions of the genome that are accessible to TF binding. We show that motifs located in these DHS sites are predictive of cell-type specific expression.

Some of the predictive motifs we identified were found to be enriched within cell-type specific DHS in a previous study using a subset of the cell types used here (Song et al. 2011). Patterns of co-occurrence and conservation of TFBS have also been used to identify regulatory modules *de novo* (Aerts et al. 2003; Sharan et al. 2003; Fu et al. 2004; Gotea and Ovcharenko 2008). However, our approach differs from such motif finding approaches, as it is not based on the sole presence of motifs, but their predictive value for gene expression patterns. As a result, the regression coefficients in our classifier and the expression profile of the TF can be viewed as testable predictions of the activating or repressing nature of the regulatory interactions between TFs and the different patterns. We also do not restrict our analysis here to cell-type specific DHS to allow for the possibility that a motif could be present in a region of ubiquitously open chromatin, but only be predictive of gene expression in a specific cell type, for instance, due to the cell-type specific expression of the factor binding to it.

CpG islands are hallmarks of unmethylated regions in vertebrate genomes and are known to overlap promoter regions, in particular in constitutively expressed genes (Yamashita et al. 2005; Carninci et al. 2006). Our results here are in agreement with previous findings that normalized CG dinucleotide content is negatively correlated with the specificity of gene expression. Consequently, constitutively expressed genes show higher CG content than up-regulated genes. While this feature is therefore useful in differentiating constitutively expressed genes, it is a confounding feature of proximal promoters when defining tissue specific regulatory codes. Our models are based on normalized binding site scores in a compendium of proximal and distal regulatory regions, and thus show consistent performance across different expression patterns. The classification performance we achieved when using the presence of motifs from open chromatin regions is significantly better than using the proximal promoter region. This is the case even when CG content is included as a feature for the classifier. We note that while using conserved regions of the genome improved the performance of the classifier over that achieved with the proximal promoter region, scanning for motifs in open chromatin regions still provided the best performance. Interestingly, in a previous study, only 43% of DHS were found to overlap an evolutionarily conserved region (Boyle et al. 2008a) and it is known that functional enhancers are sometimes weakly conserved (Blow et al. 2010).

A related recent study monitored expression using transient transfection assays for several promoters (Landolin et al. 2010). The authors then used sequence features in

the transfected plasmids to predict expression with high accuracy. There are two main differences between this work and our study reported here. First, as pointed out by Landolin et al., the promoters are not in their endogenous context in the plasmid. Therefore, this effort reflects the role sequence plays in determining expression outside the chromatin context. In contrast, our work attempts to identify cell-type specific expression from the endogenous accessibility of putative *cis*-regulatory regions. Second, the authors defined classification tasks different from the ones we examined here. In particular, they discriminated cell-type specifically expressed genes from ubiquitously expressed and unexpressed genes. While the first classification task is similar to the UR vs. constitutive classifiers here, we do not attempt to define ubiquitously unexpressed genes, as genes could always be expressed in another condition not assayed, or be affected by artifacts such as ineffective probes or mis-annotated genes. On the other hand, we build classifiers for the harder problem of predicting up- or down-regulation in one cell type versus another.

As expected, we observed an increased performance when using the comprehensive set of all PWM scores rather than just those for the most specifically expressed TFs. However, these models are harder to interpret: many PWMs used to compute the comprehensive feature vectors are highly similar or identical. TFs with the same protein binding domains also have similar binding preferences, and a large proportion of the TFs in the current release of the UniProbe database are homeodomain TFs. This can lead to collinearity among the features that are used to classify the genes

into different sets. As a result, over multiple iterations of the cross-validation, the weights assigned to each PWM are distributed to similar PWMs, and comparatively few PWMs had significant regression coefficients. To counter this, we used the subset of specifically expressed TFs, where our modeling approach allowed us to identify several known TFs that regulate gene expression but also additional candidates to study for their potential role in gene regulation in the given cell type. Future efforts will make use of recent sparse regression models that explicitly account for feature redundancy, or use projection methods such as factor analysis to explain the high-dimensional feature vectors by smaller numbers of covariates.

DNase data also showed footprints of cell-type specific binding of some factors at a high resolution. This analysis therefore corroborates recent analyses that demonstrated that the DNase-seq assay improves the signal-to-noise when attempting to identify functional locations of TF binding (Boyle et al. 2011; Pique-Regi et al. 2011). Future work in predicting gene expression will attempt to understand the utility of this high-resolution data in predicting gene expression.

4. A fine time course expression analysis identifies transcriptional cascades and a putative regulator of sex determination

4.1 Summary

The commitment of a bipotential gonad to differentiate as a testis or an ovary is governed by a dynamic transcription network that remains to be elucidated. We profiled genome-wide gene expression at <5hr resolution during the critical one-day developmental window in XX and XY mouse gonads from the 129S1/SvImJ and C57BL/6J strains. We identified cascades of expression in both strains and show that establishment of dimorphic expression is largely due to activation and repression programs initiated by the testis pathway. Strikingly, comparison of expression differences between the two strains revealed a delay in the male program in C57BL/6J gonads, suggesting an explanation for the increased susceptibility to male-to-female sex reversal exhibited by this strain. Finally, we exploit the predictive power inherent in this temporal dataset to identify a novel candidate gene, *Lmo4,* underlying an expression QTL identified in a previous study. Confirming our prediction, knockdown of this gene in primary XY gonad cells resulted in the down-regulation of several male program genes. Our results highlight the importance of fine-scale resolution time-course measurement of expression in developmental systems to identify candidate regulatory genes and to understand general properties of the system.

4.2 Introduction

To improve our resolution of the transcriptional cascade controlling sex determination, and choose attractive candidates in eQTL intervals, we conducted a fine time course transcriptome analysis of the gonad between E11.0-E12.0, when the bipotential gonad approaches a decision point, initiates the testicular or ovarian pathway, and begins to reinforce the sexual fate decision. We profiled global gene expression at six equally-spaced intervals in XX and XY gonads from the susceptible B6 and resistant 129S1 strains, developed and trained a Hidden Markov Model (HMM) to discern the onset of sexually-dimorphic expression, and identified gene cohorts activated or repressed specifically in the testis or ovary during this brief 24-hour window of development. Specifically, we assayed total transcript abundance in XY and XX gonads at six equally spaced intervals between E11.0-E12.0 (Figure 14). By comparing the onset profiles of both strains, we found that susceptibility to sex reversal in B6 XY gonads is likely due to the delayed activation of many testis pathway genes and delayed repression of many ovarian pathway genes. We exploited this detailed view of the B6 and 129S1 gonad transcriptomes to prioritize candidate regulatory genes. Finally, we developed a primary cell validation assay and lentivirus-based shRNA delivery method to artificially silence *Lmo4* (Lim-domain only 4), a candidate regulatory gene within an eQTL interval. We provide strong evidence that *Lmo4* is a novel regulator of sex determination upstream of many sex-associated genes. This work provides a

systematic framework for predicting and testing regulatory genes (eQTGs) underlying eQTLs that is applicable to other systems.

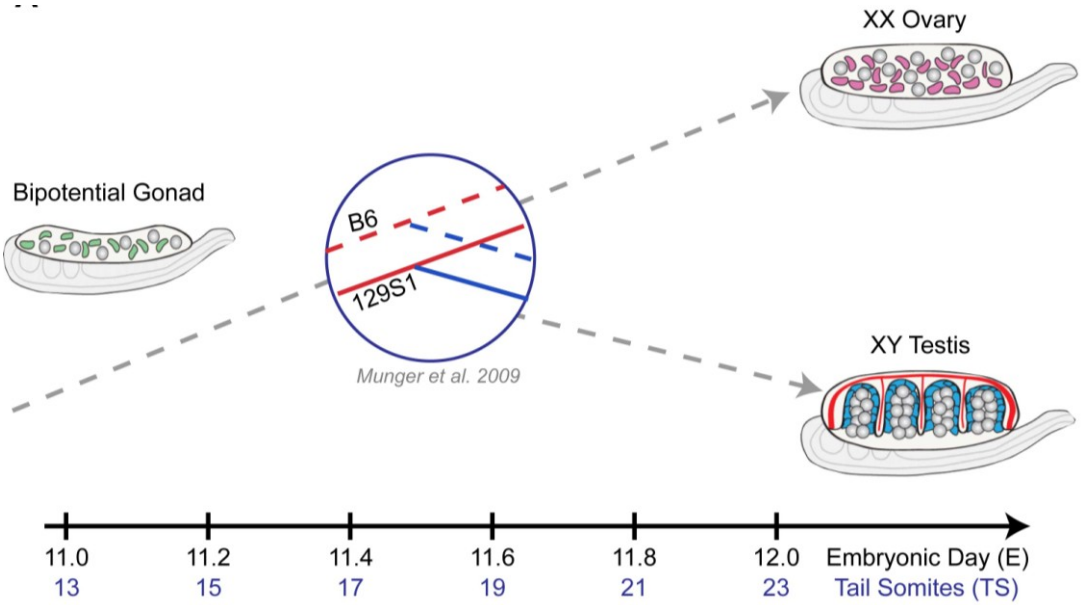


Figure 14: Experimental design for profiling the gonad transcriptome during the 24-hour period encompassing sex determination.

Total transcript abundance was profiled in XX and XY gonads at six equally spaced intervals between E11.0-E12.0 capturing the critical transition in the gonad transcriptome from a bipotential to sexually-differentiated state. The analysis was conducted in two inbred strains, C57Bl/6J (B6) and 129S1/SvImJ (129S1), with well-characterized differences in their sensitivity to sex reversal.

4.3 Results

4.3.1 Dynamic transcriptional changes are observed as the bipotential gonad differentiates to a testis and ovary

Total transcript abundance was measured by microarray for individual pairs of gonads for each sex/strain/stage combination (n = 74 total arrays). A total of 9,254 genes (12,213 probes) exhibited significant expression above background in at least two replicates of one sample type, and were included in subsequent analyses (Figure 15). Next, we fit a linear model accounting for the effects of strain, sex, stage, and two-way (e.g. sex*stage) and three-way (e.g. sex*stage*strain) interactions among these factors (Figure 16). For more than half (n = 4,752 (5,659 probes)) of the genes that passed our filtering criteria, a significant proportion of the observed variation in expression could be attributed to one or more experimental variables (Figure 15). The individual components of sex (n = 1,172 genes/ 1,343 probes), stage (n = 2,434 genes/ 2,805 probes), and strain (n = 3,279 genes/ 3,879 probes), as well as the interaction effect of sex by stage (n = 659 genes/ 733 probes), all had significant effects on the expression of hundreds to thousands of genes. For many of these genes, expression was influenced by the additive effects of multiple components (probes in overlap regions in Venn diagram, Figure 15). Moreover, the sex by stage interaction effect reflects the number of genes that have a sexually-dimorphic pattern of expression that changes over time in the E11.0-E12.0 window. Finally, the large number of genes with a strain effect highlights the extent to which the transcription programs vary in the B6 and 129S1 strains. These data illustrate

both the complexity and dynamic nature of the transcriptional program driving sex determination during this brief but critical developmental window.

QC and ANOVA		# Probes	(genes)
Total probes on array		**25697**	(18138)
Probes expressed above background		**12213**	(9254)
Probes with at least one significant component		**5659**	(4752)
ANOVA Results	Sex	**1343**	(1172)
	Stage	**2805**	(2434)
	Strain	**3879**	(3297)
	Sex*Stage	**733**	(659)
	Sex*Strain	**35**	(34)
	Stage*Strain	**41**	(41)
	Sex*Stage*Strain	**11**	(11)

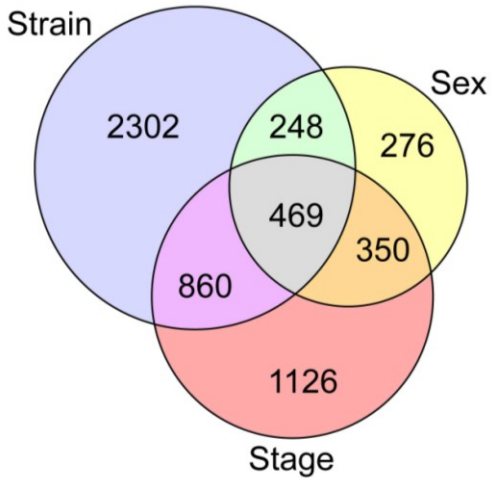


Figure 15: Analysis of Variance (ANOVA) identifies genes showing significant variation by strain, sex, stage and their interactions

Nearly half of the probes on the array (n = 12,213), representing more than half of all genes (n = 9,254, shown in gray), were expressed above background levels at one or more time points between E11.0-E12.0 in XY or XX gonad samples. For a large proportion of expressed probes (n = 5,659), variation in gonad transcript abundance was significantly associated with additive effects from sex, developmental stage, and/or strain. A sex-by-stage interaction effect accounted for a significant proportion of the expression variation in 733 probes. For many probes, variation in expression is driven by more than one experimental variable. A Venn diagram (Right) showing probes whose expression is affected by one or more of the additive effects of sex, stage, and strain. Values within each circle correspond to the number of probes that are significantly affected by that variable. Note that the overlaps in the Venn diagram do not

capture interaction effects, but represent probes that are significantly affected by two or all three factors.

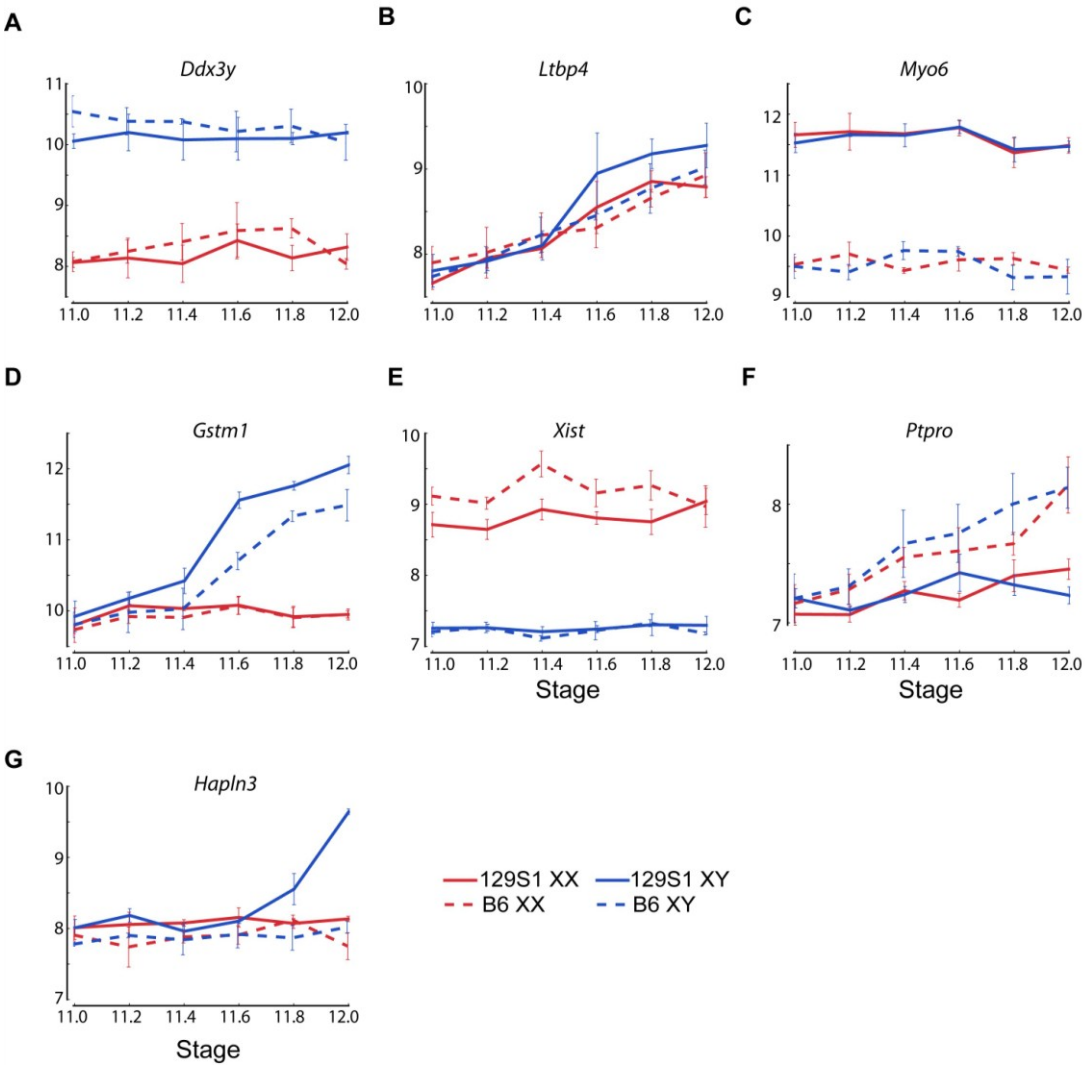


Figure 16: Examples showing significant difference in expression for each of the variables in the ANOVA analysis.

(A) Sex effect (*Ddx3y*): XY gonads show higher expression than XX gonads. **(B)** Stage effect (*Ltbp4*): Expression is higher at later time points across strains and sex. **(C)** Strain effect (*Myo6*): 129S1 mice show higher expression regardless of sex and stage. **(D)** Sex-by-stage effect (*Gstm1*): XY gonads show higher expression starting at E11.6. **(E)** Sex-by-strain effect (*Xist*): B6 XX gonads show higher expression compared to 129S1 XX

gonads. **(F)** Stage-by-strain effect (*Ptpro*): B6 gonads show higher expression at later stages. **(G)** Sex-by-stage-strain (*Hapln3*): 129S1 XY gonads starting at E11.8 show significantly different expression compared to all other sample types.

4.3.2 Developing a Hidden Markov Model to identify cascades of expression

To identify ordered cascades of expression and co-regulated genes, we developed a Hidden Markov Model (HMM) (Figure 17). HMMs are well-suited to the task of discerning patterns in time series data (Schliep et al. 2003; Yuan and Kendziorski 2006) because they use correlations between adjacent time points to overcome noise and increase sensitivity. Briefly, the HMM was designed with 18 states, three per time point – a male (i.e. testis-enriched) state, a female (i.e. ovary-enriched) state, and a similar expression state (Figure 17, Figure 18). The fold difference of a gene's expression between XX and XY gonads at each time point was used to train the HMM. After training the model, the Viterbi state path of each gene reflected whether the gene was expressed in a sexually dimorphic fashion, the sex in which it was expressed more highly, and the times at which the gene exhibited dimorphic expression. Importantly, only 22 of a possible 729 state paths through the model were populated (Table 7), indicating that despite the highly dynamic changes in the transcriptome, there are common expression trajectories by which the expression patterns can be clustered. We note that while the HMM classifies genes as becoming dimorphically expressed at specific times, this is due to the discrete sampling during the window. In reality, these genes are likely to show minor deviation from the specific time points at which we identify them as becoming dimorphic. Nonetheless, grouping the genes by their onset of dimorphism reveals interesting details of the regulatory programs involved.

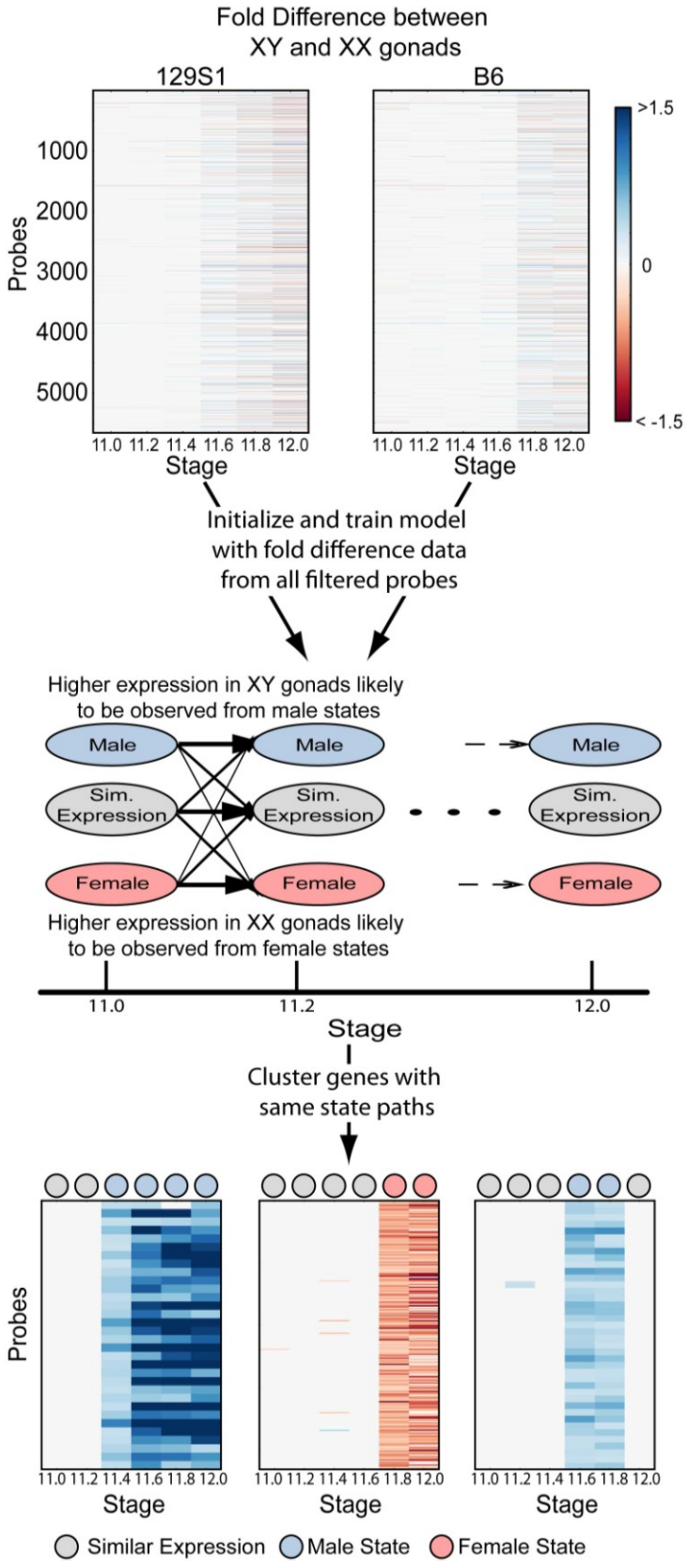

Fold Difference between
XY and XX gonads
129S1
B6
Probes
1000
2000
3000
4000
5000
>1.5
0
< -1.5
11.0 11.2 11.4 11.6 11.8 12.0
Stage
Initialize and train model
with fold difference data
from all filtered probes
Higher expression in XY gonads likely
to be observed from male states
Male
Sim.
Expression
Female
Higher expression in XX gonads likely
to be observed from female states
11.0
11.2
12.0
Stage
Cluster genes with
same state paths
Similar Expression
Male State
Female State

Figure 17: A Hidden Markov Model (HMM) to identify patterns of dimorphic expression in the gonad transcriptome.

Fold differences between XY and XX gonads at each time point in both strains were calculated for all probes passing the ANOVA filtering step. This data was then used to initialize and train the Hidden Markov Model (HMM) (see Methods). The most probable (Viterbi) state path reflects possible dimorphic expression patterns between XX and XY gonads and was used to cluster genes. Heatmaps illustrate 3 clusters with state paths indicated by circles at the top of each heatmap.

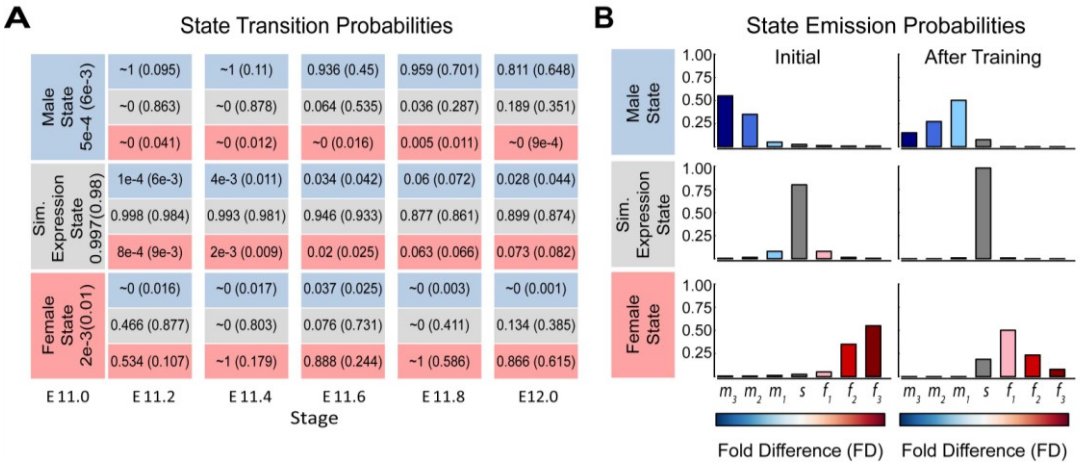


Figure 18: State transition and emission probabilities of the Hidden Markov Model before and after training

(A) State transition probabilities for the HMM (Figure 17) after (before) training with the Baum-Welch algorithm. Numbers in the E11.0 column show probability of a gene's expression starting in the male, female or similar expression state after (before) training. Transition probabilities are shown for each pair of transitions from one time point to the next for E11.2-E12.0. Colors of the cells indicate the state at that time point. First three rows in each column show transition from a male state at the previous time point, the middle three show transition from a similar expression state and the last three rows show transition from a female state at the previous time point. For example, after training, the probability of transitioning from a similar expression state at E11.6 to a male expression state at E11.8 is 0.06. **(B)** State emission probabilities for the three states

before (left panels) and after training (right panels). Emission probabilities for discretized fold changes were initialized by hand. After training, emission probabilities still reflect the intuitive meaning of the states. For example, higher expression in XY gonads is likely to be observed in emissions from the male state. Emission probabilities for the states were tied across time points.

Table 7: Numbers of genes (probes) in each Viterbi state path identified by the HMM for both 129S1 and B6 mice.

	Viterbi State Path	Number of Genes (Probes) in 129S1	Number of Genes (Probes) in B6
Male-Enriched Genes	○●●●●●	1(1)	1(1)
	○○●●●●	30(32)	14(14)
	○○○●●●	191(208)	71(75)
	○○○○●●	141(149)	260(275)
	○○○○○●	202(208)	159(170)
Female-Enriched Genes	○●●●●●	7(7)	0(0)
	○○●●●●	15(16)	1(1)
	○○○●●●	148(167)	24(25)
	○○○○●●	200(216)	191(215)
	○○○○○●	378(395)	312(333)
	○○●●●○	2(2)	1(1)
	○○○●●○	40(40)	16(16)
	○○●○○○	1(2)	0(0)
	○○○●○○	10(10)	2(2)
	○○○○●○	8(8)	17(17)
	●○○○○○	0(0)	1(1)
	○●●○○○	2(2)	0(0)
	●●●●●●	3(3)	3(3)
	●●●●●●	3(5)	5(7)
	○●●●●●	1(1)	0(0)
	○○○●●●	2(2)	0(0)
	○○○○○○	3587(4185)	3868(4503)

○ Similar Expression ● Male State ● Female State

Each state path has six states, one for each time point, and genes having the same state path are clustered together. Only 3 genes switch from showing higher expression in one sex to higher expression in the other. Most genes that become dimorphic continue showing dimorphism throughout the E11.0–E12.0 window.

4.3.3 Identifying sequential cascades of expression by onset of dimorphism

Out of the 4,752 genes included in the analysis, 1,321 genes exhibited dimorphic expression at one or more time points between E11.0-12.0 in the 129S1 strain and similar numbers (1,037 genes) were dimorphically expressed in the B6 strain (Table 7). Interestingly, for both 129S1 and B6, once a gene established a dimorphic expression pattern, most continued in a state of sexually dimorphic expression until E12.0 (n = 1,254 genes for 129S1, n = 995 genes for B6). We refer to these genes as male- or female-enriched depending on which sex exhibited higher expression. Finally, only three genes (*Lefty2*, *Mcm6*, and *LOC233529*) in 129S1 (and none in B6) switched from being more highly expressed in one sex to the other during the duration of our window.

We used the HMM to cluster male- and female-enriched genes by the time of onset of dimorphic expression from E11.2 to E12.0 for the 129S1 strain (Figure 19). This analysis revealed striking cascades of sexually-dimorphic male and female enrichment with the number of male- and female-enriched genes gradually increasing across time points. For example, for male-enriched genes a single gene (*Sox9*) showed higher expression in males starting at E11.2, followed by 30 genes at E11.4, and finally 202 genes that showed sexually dimorphic expression at E12.0 (Table 7). To determine whether these cascades primarily reflected changes in one gonadal cell type or several, we compared our whole gonad data with cell type-specific gene expression data from E11.5 and E12.5 isolated XX and XY supporting cells and germ cells (Jameson et al.

2012b). We found that the overlap with germ cells was low (5%). In contrast, 58% of genes that became male- or female- enriched in our whole gonad transcriptome prior to E11.8 were specifically dimorphic at E11.5 or E12.5 in supporting cells. After E11.8, the overlap with the supporting cell precursors dropped to 45%. Thus, consistent with previous results, the supporting cell lineage, known to be critical for initiating the sex determination decision, is responsible for a large proportion of the sexually-dimorphic gene expression that arises in the gonad between E11.0-E12.0 (Jameson et al. 2012b).

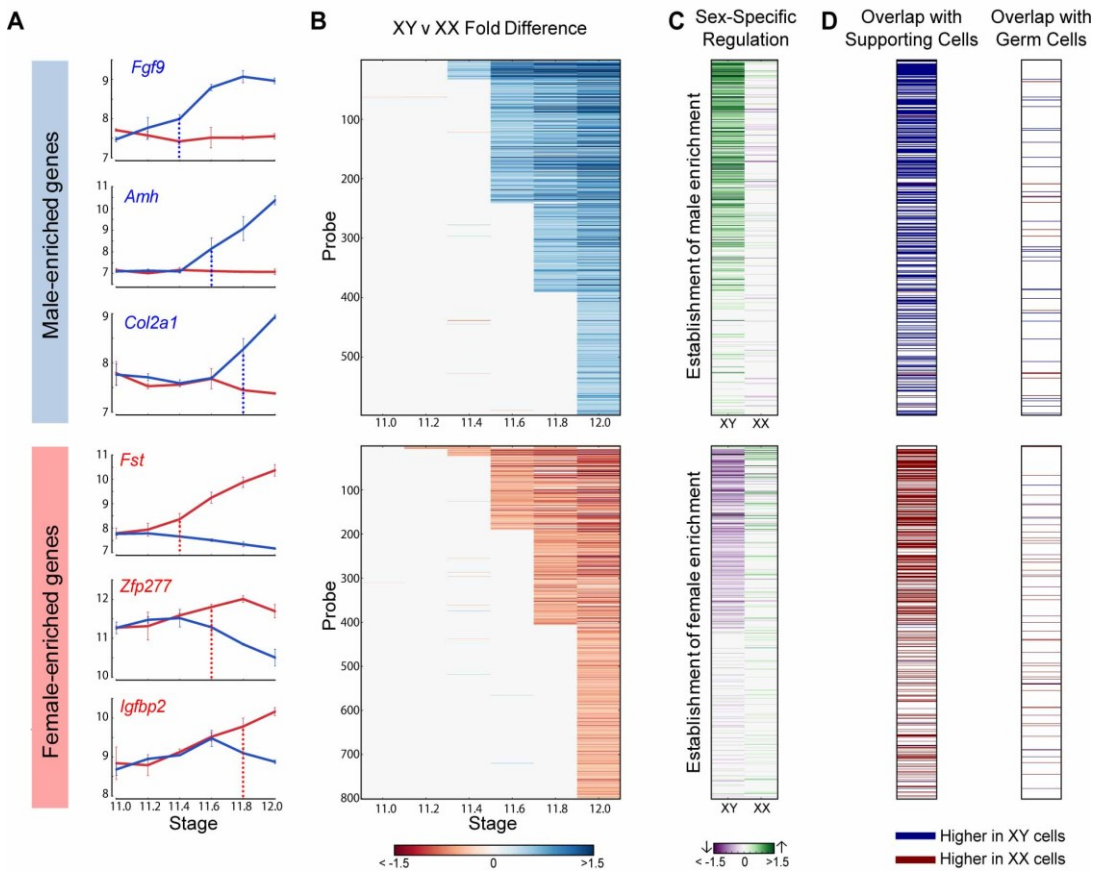


Figure 19: Cascades of dimorphic expression involving both activation and repression in XY gonads.

(A) Examples of genes showing higher expression in XY (male-enriched genes, top panel) and XX gonads (female-enriched genes, bottom panel) from 129S1 mice. Blue and red vertical lines show the time of onset of dimorphic expression. **(B)** Cascades of dimorphic gene expression identified by the HMM in XY (top panel) and XX gonads (bottom panel). Colors indicate the log fold change between XY and XX gonads at a specific time point for the 129S1 strain. The genes are arranged in order of time of onset of dimorphic expression. **(C)** Contribution to changes in expression between E12.0 and the time point before the onset of dimorphism are shown for each gene in (B) in XY

(column 1) and XX (column 2) gonads. Top panel: male-enriched genes. Bottom panel: female-enriched genes. This analysis shows that male-enriched genes are mostly up-regulated in XY gonads while female-enriched genes are mostly down-regulated in XY gonads. **(D)** Male- or female-enriched genes in whole gonads between E11.0 and E12.0 are primarily dimorphic in the supporting cell lineage of the gonad in 129S1 mice. The cascade of genes that become dimorphically expressed between E11.2 and E12.0 was cross-referenced with cell-type specific expression datasets analyzed at E11.5 and E12.5 (Jameson et al. 2012b). Column 1 shows overlap with genes expressed dimorphically in supporting cells while column 2 shows overlap with genes expressed dimorphically in germ cells. Rows are colored blue or red where the probe was dimorphically expressed and higher in XY cells or higher in XX cells, respectively. The highest overlap is seen with the supporting cells for both male- and female-enriched genes.

The HMM classified *Sox9* as the earliest male-enriched autosomal gene (E11.2), reflecting its position directly downstream of *Sry* in the testis pathway (Sekido and Lovell-Badge 2008) and affirming the fine temporal resolution of our dataset. As in previous microarray experiments using whole gonad samples (Small et al. 2005), *Sry* was not detected above background levels in the current study. Following the up-regulation of *Sox9*, several other known crucial downstream genes such as *Fgf9*, *Amh*, and *Dhh* showed increased expression in XY gonads (Figure 19, top panel). In the female-enriched group, *Wnt4*, one of the earliest autosomal genes known to act in ovary differentiation (Vainio et al. 1999; Jameson et al. 2012a), was sexually-dimorphic at E11.4. Other known female pathway genes such as *Fst* and *Axin2* became differentially expressed at the same stage or immediately following the dimorphic expression of *Wnt4*. All the male-enriched genes that were activated prior to *Sox9* and female-enriched genes enriched prior to *Wnt4* are Y- and X-linked genes, respectively (Figure 20 and Figure 21). Particularly interesting are the 7 X-linked genes that exhibited higher expression in XX gonads starting at E11.2 (Figure 21). The cell-type specific data indicate that these genes are all highly expressed in germ cells and likely reflect the reactivation of the inactive X chromosome in XX germ cells at this stage (Chuva de Sousa Lopes et al. 2008).

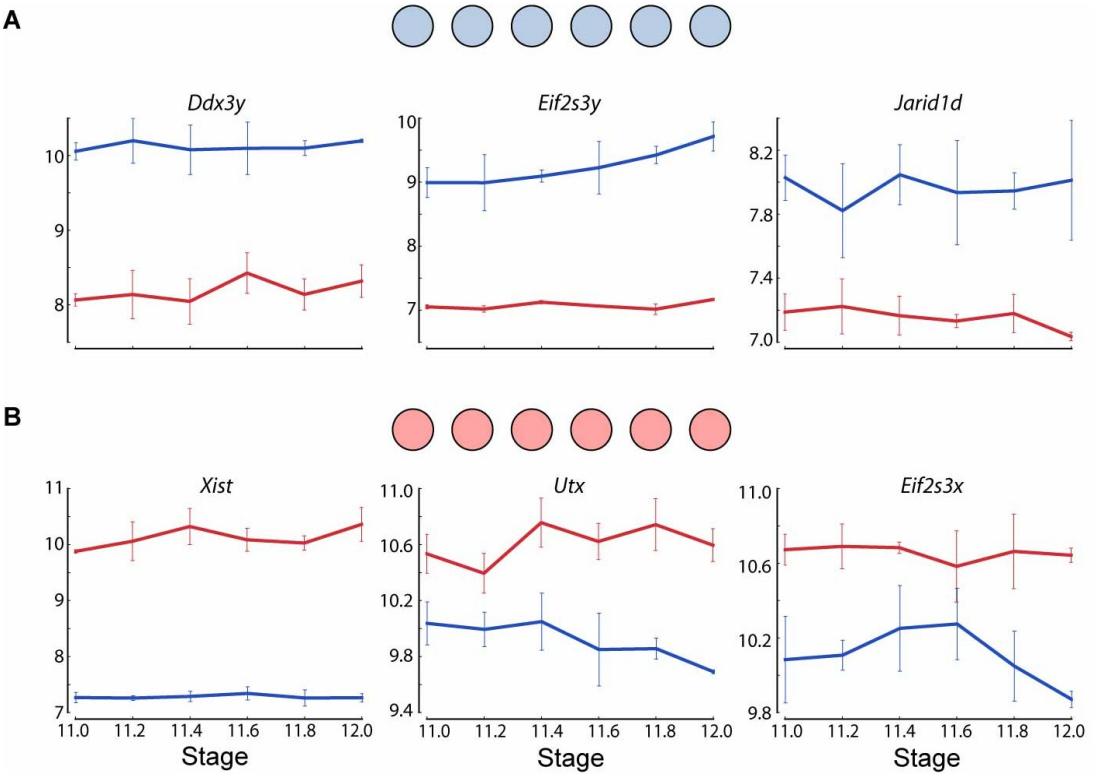


Figure 20: X-linked and Y-linked genes that are dimorphically expressed across the E11.0 – E12.0 window.

(A) Expression of *Ddx3y*, *Eif2s3y,* and *Jarid1d* in 129S1 gonads (blue – expression in XY gonads, red – expression in XX gonads). All three genes are Y-linked and are expressed higher in XY gonads across the E11.0 – E12.0 window. **(B)** Expression of *Xist*, *Utx*, and *Eif2s3x* in 129S1 gonads. All three genes are X-linked and are expressed higher in XX gonads across the E11.0 – E12.0 window.

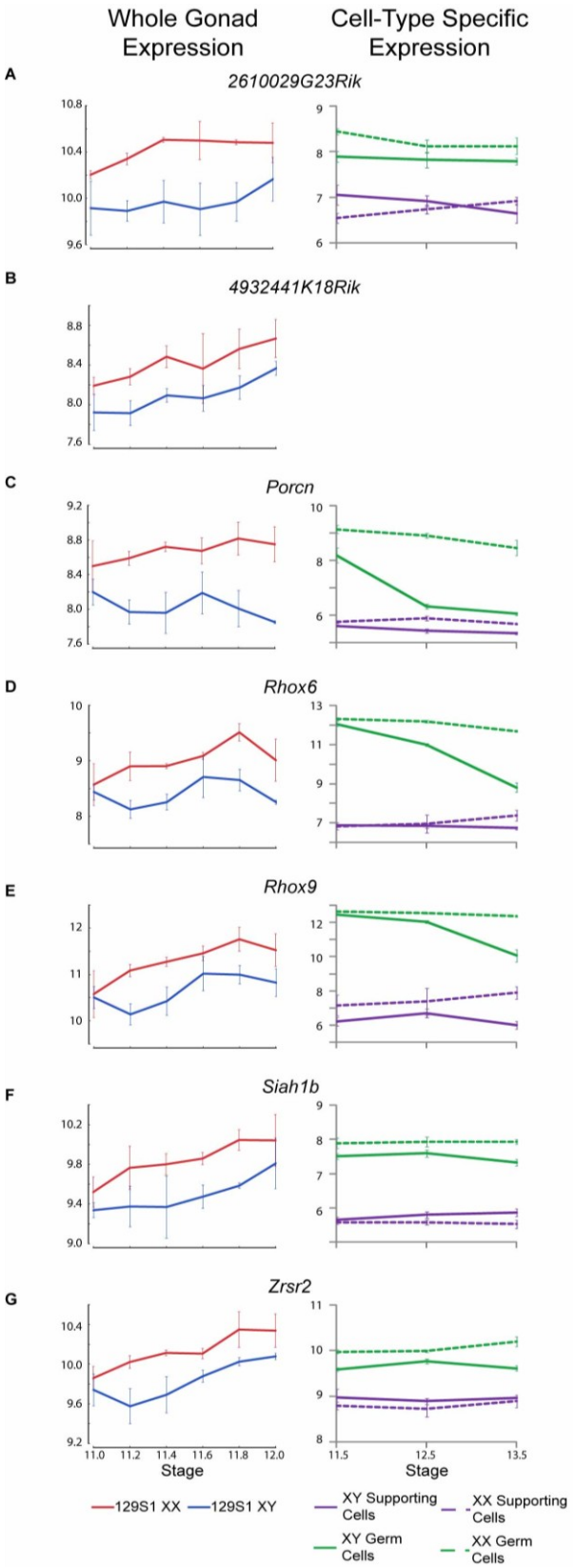

Whole Gonad Expression
Cell-Type Specific Expression
A
2610029G23Rik
B
4932441K18Rik
C
Porcn
D
Rhox6
E
Rhox9
F
Siah1b
G
Zrsr2
Stage
Stage
129S1 XX
129S1 XY
XY Supporting Cells
XX Supporting Cells
XY Germ Cells
XX Germ Cells

Figure 21: X-linked genes showing higher expression starting at E11.2 are germ-cell enriched.

(A) – (G). (left column) 7 genes showing higher expression in XX gonads in 129S1 mice, likely due to the higher number of germ cells in 129S1 mice (Western et al. 2011). (**A) – (G)**. (right column) Corresponding expression in male (solid line) and female (broken line) in germ cells (green) and supporting cells (purple) from cell-type specific expression data. Note that *4932441K18Rik* expression was not captured in the cell-type specific expression dataset. All genes show enriched expression in germ cells.

4.3.4 The testis program activates and represses large cohorts of genes

To determine whether activation, repression, or a combination of both was involved in primary establishment of dimorphism for each gene, we conducted an analysis using the initial and final time points in our time course. We calculated the fold change in both XX and XY gonads for each gene between E11.0 and E12.0 in the robust 129S1 strain. We then graphed these sex-specific fold changes for each gene on an X-Y scatter plot, with fold changes in the XY gonad appearing on the Y-axis, and changes in the XX gonad appearing on the X-axis (Figure 22). The scatter plot identified genes exhibiting male enrichment (*Sox9*, *Dhh,* and *Cbln1* (Bitgood et al. 1996; Morais da Silva et al. 1996)), or female enrichment (*Irx3, Wnt4,* and *Msx1* (Menke and Page 2002; Jorgensen and Gao 2005)), are shown adjacent to their locations in the scatter plot. In addition to genes with sexually dimorphic expression patterns, we also identified genes that are identically-repressed or activated by both the male and female programs (genes on the diagonal in Figure 22). We hypothesize that identically-activated genes (e.g. *Hoxd10* and *Hs3st3a1*) canalize sexual differentiation regardless of the nature of the fate commitment, while genes that are identically-repressed (e.g. *Gfra3* or *Hoxa7*) function during early gonad formation to preserve the bipotential state of the gonad. A subset of these genes (14, log2(fold change) >0.585 in both sexes) overlaps with the "core adrenogonadal program" previously identified in the related steroidogenic-factor-1-positive cell population (Pitetti et al. 2013).

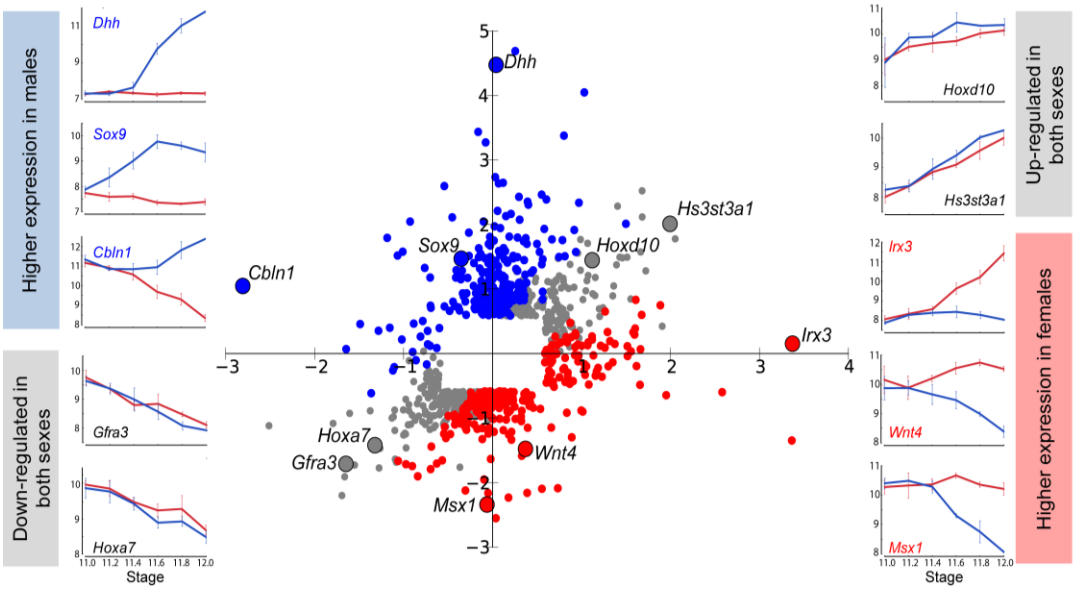


Figure 22: Changes in XX and XY gonads contribute to expression fold change between E11.0 and E12.0.

Gene expression in XY and XX gonads was compared at the beginning and end of the 24-hour developmental window. For probes that exhibited a 1.5-fold or greater change in expression in either sex between E11.0 and E12.0, log of the Fold Change in the XY gonad is plotted on the Y-axis, and log of the Fold Change in the XX gonad is plotted on the X-axis. Probes that are similarly up-regulated or down-regulated in both sexes appear in gray in the upper right and lower left quadrants, respectively. Probes that become enriched in XY gonads relative to XX are shown in blue, while genes that become enriched in XX gonads relative to XX are shown in red. Examples from each category are highlighted, and their expression patterns in XY (blue line) and XX (red line) gonads are displayed.

This view of sexually dimorphic expression changes between E11.0 and E12.0 also revealed that higher expression in one sex can result from activation in one sex (e.g. *Dhh* in Figure 22), repression in the other sex (e.g. *Msx1*), or a combination of both mechanisms (e.g. *Wnt4*). From the scatter plot, it is evident that dimorphic expression of most genes (202 genes, log2(fold change) > 0.585) expressed higher in the testis occurred primarily through activation, with a small outlier group of genes (24 genes, log2(fold change) < -0.585) showing dimorphism as the result of repression in the XX gonad. Among genes showing higher expression in XX gonads, two principal gene clusters were evident: members of one cluster (76 genes) achieved dimorphism primarily through activation in the XX gonad, and members of the other (147 genes), primarily through repression in the XY gonad. This indicates that the dynamic expression changes observed during gonad fate commitment are a result of the action of activation and repression programs.

Note that the analysis of the scatter plot does not take into account changes in expression initiated at the onset of dimorphism. To do so, we compared gene expression in each sex before the initiation of differential expression and the E12.0 stage (Figure 19C). For all genes that showed higher expression in the XY gonad by E11.8, 73% of genes (256 genes) were strongly activated (log2(fold change) > 0.32, $p < 0.05$) in XY gonads, whereas only 9.5% (34 genes) were repressed in the XX gonad (log2(fold change) < -0.32, $p < 0.05$). In addition 5.6% (20 genes) become dimorphic through a

combination of activation in XY and repression in XX gonad. This indicates a strong activation program in XY gonads with a much lower contribution from repression of male pathway genes in XX gonads. In striking contrast, enrichment of genes in XX gonads results not from activation in the ovary, but primarily through repression in the testis (Figure 19C, lower panel). Only 16% of probes (61 genes) that are female-enriched by E11.8 are activated in the XX gonad, while 61% (217 genes) are repressed in the XY gonad, with 7.5% (27 genes) becoming dimorphic due to a combination of activation in the XX and repression in the XY gonad. In fact, in several cases, *Msx1* for example (Figure 22), female enrichment stems exclusively from repression taking place in the testis. This indicates that in addition to the activation program, a strong repressive program is also present in the testis. While male repression of specific female pathway genes (*Wnt4*) has been known, the extent of this repressive signature is surprising.

4.3.5 Activation of the male and repression of the female differentiation pathways in the XY gonad are delayed in the sensitive B6 strain

After characterizing the temporal dynamics of XY and XX gonad transcriptomes from the 129S1 strain that is resistant to XY sex reversal, we determined how the transcriptome varied in B6, a strain that is sensitive to XY sex reversal in response to multiple genetic perturbations (Eicher et al. 1982; Bouma et al. 2005; Correa et al. 2012). While previous studies showed that male-enriched genes were expressed at a higher level and female-enriched genes at a lower level in 129S1 compared to B6 E11.5 XY

gonads (Munger et al. 2009), it was not clear whether this strain difference was a result of the difference in expression levels, time of onset, or a combination of both. To address these questions, we profiled global gene expression in XY and XX gonads from B6 at the same six time points spanning the critical 24-hour window (E11.0-E12.0). We then used the HMM to identify male- and female-enriched genes in B6 mice (Figure 23). In good agreement with the data from 129S1 mice, these genes showed activation and repressive programs in XY gonads, and were strongly biased toward dimorphic expression in supporting cells.

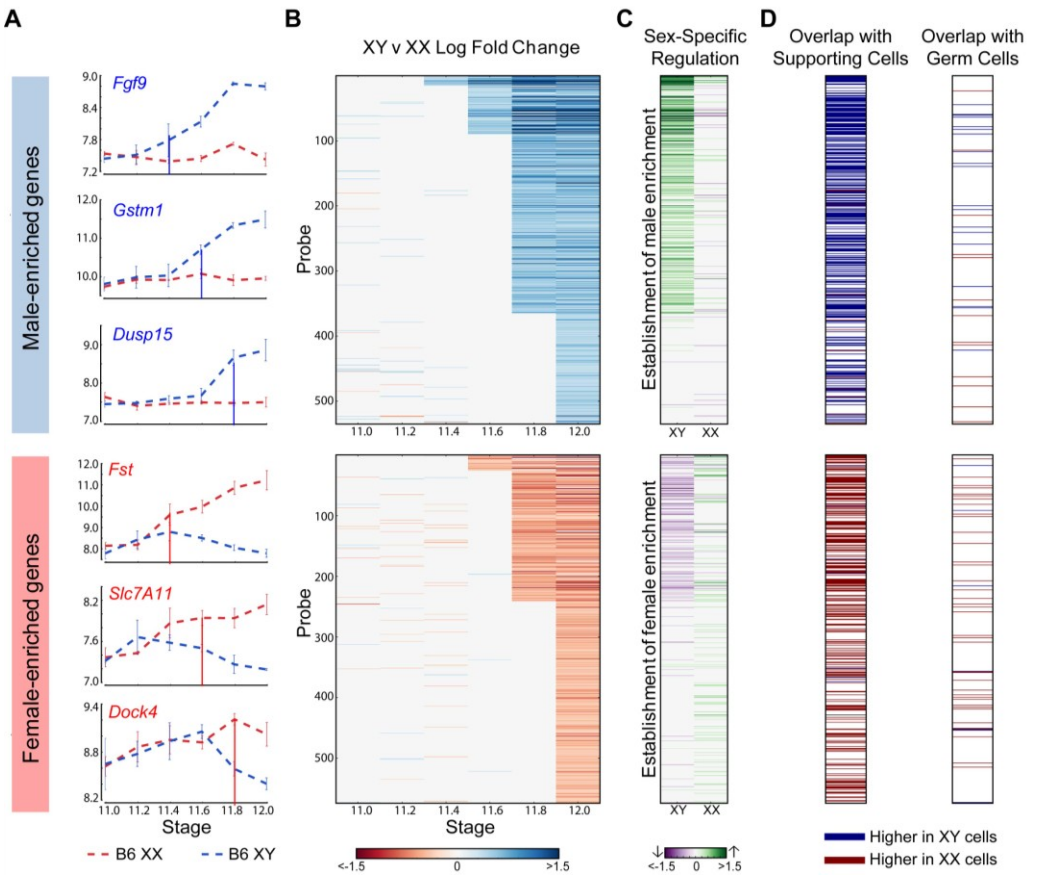


Figure 23: Detailed characterization of dimorphic expression in B6 gonads reveals properties similar to 129S1 gonads.

(**A**) Examples of genes showing higher expression in XY (male-enriched genes, top panel) and XX gonads (female-enriched genes, bottom panel) from B6 mice. Blue and red vertical lines show the time of onset of dimorphic expression. (**B**) Cascades of dimorphic gene expression identified by the HMM in XY (top panel) and XX gonads (bottom panel). Colors indicate the fold difference between B6 XY and XX gonads at a specific time point. The genes are arranged in order of increasing time of onset of dimorphic expression. (**C**) Contribution to changes in expression between E12.0 and the time point before the onset of dimorphism are shown for each gene in (B) in XY (column

1) and XX (column 2) gonads. Top panel: male-enriched genes. Bottom panel: female-enriched genes. This analysis shows that male-enriched genes are mostly up-regulated in XY gonads while female-enriched genes are mostly down-regulated in XY gonads. (**D**) The cascade of genes dimorphically expressed was cross-referenced with cell-type specific expression datasets analyzed at E11.5 and E12.5 (Jameson et al. 2012b). Column 1 shows overlap with genes expressed dimorphically in supporting cells while column 2 shows overlap with genes expressed dimorphically in germ cells. Rows are colored blue or red where the probe was dimorphically expressed and higher in XY cells or higher in XX cells, respectively. As with 129S1 gonads, the highest overlap is seen with the supporting cells for both male- and female-enriched genes in B6 gonads.

We then compared the timing of onset of sexually dimorphic gene expression between 129S1 and B6 (Figure 24). We observed a clear, consistent temporal shift in the onset of sexually-dimorphic expression in many genes in the susceptible B6 strain (Figure 24A, B). Specifically, for male-enriched genes, a comparison of strain onset distribution profiles revealed a statistically significant ~5-hour delay in B6 relative to 129S1 (Figure 24C, upper panel). For example, 208 probes became enriched in 129S1 XY gonads relative to XX starting at E11.6 (Figure 24B). When the same probes were examined in the sensitive B6 strain, only 37 became dimorphic at the same stage, while a majority (n = 107) became male-enriched ~5 hours later at E11.8. Another 20 from this set did not become dimorphic in B6 until E12.0, and 35 probes that were male-enriched in 129S1 at E11.6 failed to become sexually-dimorphic in B6 by E12.0. Even in the case of genes that became male-enriched at the same time point in B6 and 129S1, a comparison of XY v. XX fold difference at the onset of dimorphism indicated that a majority of these genes (65.6%) show a higher male v. female fold difference in 129S1 than in B6 (Figure 25A, Binomial test p-value < 0.005). This difference may reflect a more robust activation mechanism driving the male differentiation pathway in 129S1, or it could reflect a delay in expression onset in B6 that is less than ~5 hours and therefore smaller than the minimum resolution threshold of this analysis.

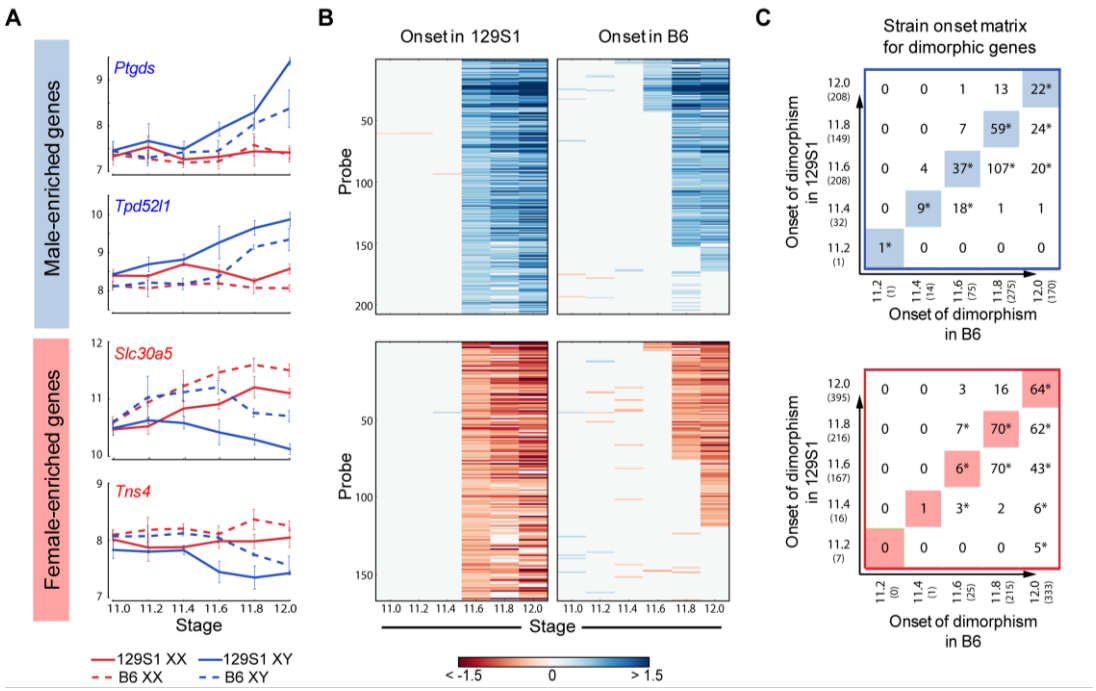


Figure 24: Dimorphic expression of multiple male- and female-enriched genes in B6 is delayed compared to 129S1 mice.

(**A**) Expression of male- (top panel) and female-enriched (bottom panel) genes. Dimorphic expression for these genes is delayed by ~5 hours in B6 compared to 129S1. (**B**) Heatmap showing dimorphic expression at E11.6 in 129S1 and comparison of same genes in B6. While a few genes show earlier dimorphic expression in B6 mice compared to 129S1, the dominant pattern shows a ~5 hr delay between B6 and 129S1 mice. (**C**) Matrix showing the time of onset of dimorphism in 129S1 and B6 mice for male-enriched (top panel) and female-enriched (bottom panel) genes. For example, out of the 32 male-enriched probes that became dimorphically expressed at E11.4 in 129S1, 9 probes were also dimorphically expressed starting at E11.4 in B6 while 18 showed dimorphic expression starting at E11.6 in B6 mice. However, 4 genes are dimorphic in B6 XY

gonads at E11.4, but not in 129S1 until E11.6. * indicates significant overlap with $p < 0.001$ evaluated by a hypergeometric test. The highlighted diagonals show the number of genes showing similar onset of dimorphism in 129S1 and B6 mice. Note that some genes that are male- or female-enriched in one strain do not show dimorphism in the other strain.

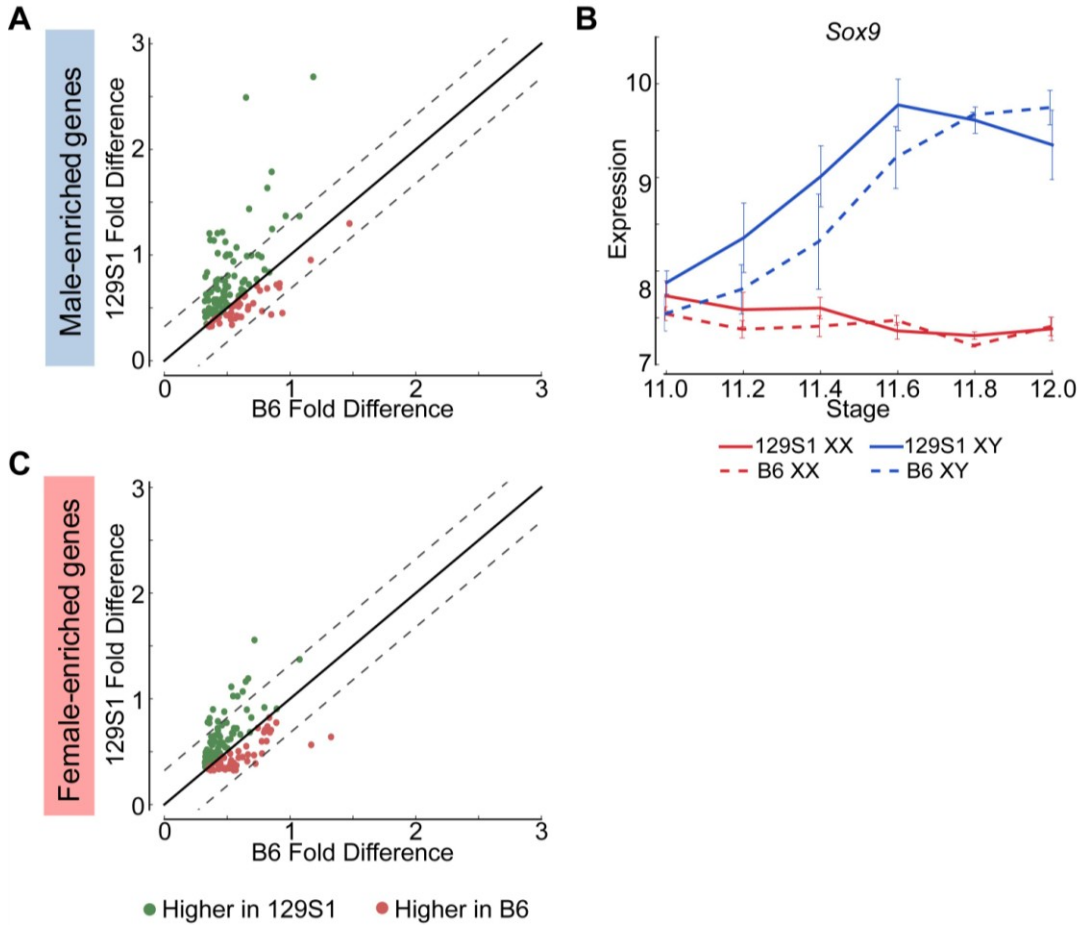


Figure 25: Robust onset of dimorphism in 129S1 mice compared to B6 mice

(A, C) Scatterplot showing XY vs. XX fold difference (A) and XX vs. XY fold difference (C) at the onset of dimorphism for male- and female-enriched genes that are activated at the same stage. Fold difference between 129S1 XY and XX gonads at the onset of dimorphism are plotted on the y-axis and the fold difference between B6 XY and XX gonads at the onset of dimorphism on the x-axis. Onset of dimorphism is more robust in the 129S1 strain for both male- and female-enriched genes. (**B**) *Sox9* becomes

dimorphic at E11.2 in 129S1 and B6 gonads. However, the fold difference between XY and XX gonads is higher at E11.2 in 129S1 mice.

Importantly, this delayed onset pattern of male-activated genes in B6 does not appear to stem from a difference at the top of the cascade in *Sry* expression level during this critical window. Although *Sry* was not detected above background in these arrays, there was no significant difference in expression levels between 129S1 and B6 between E11.2-E12.0 by qRT-PCR (Figure 26A). It should be pointed out that the high variability in *Sry* abundance observed among individual pairs of XY gonads could mask a small but real strain effect for *Sry* expression levels. In contrast, *Sox9* is more robustly up-regulated in 129S1 relative to B6 (Figure 25B), an observation we confirmed by qRT-PCR (Figure 26B). By E11.8 expression of *Sox9* in B6 XY gonads has caught up with expression in 129S1. Some of the delay in onset of male-enriched genes at later stages in B6 may be due to this initial deficiency in the robustness of *Sox9* activation.

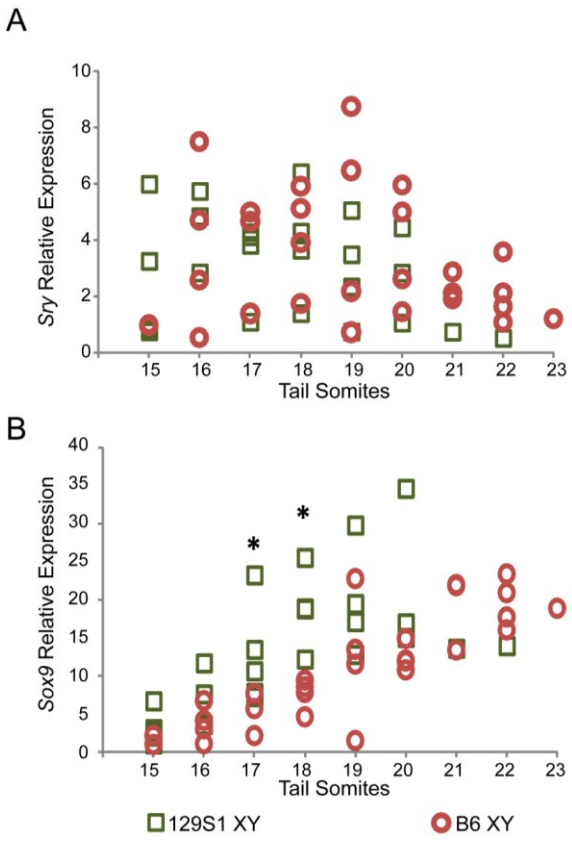


Figure 26: 129S1 and B6 XY gonads show no significant difference in *Sry* expression but a small difference in *Sox9* expression as assayed by qRT-PCR

(**A**) *Sry* expression levels are similar in 129S1 and B6 XY gonads between E11.2-E12.0. No statistically significant ($p < 0.1$) differences are detected at any time point in this analysis, however high variability among individuals may mask a small but biologically meaningful strain effect for *Sry* transcript abundance in this window. (**B**) *Sox9* expression shows significantly different expression ($p < 0.1$ (*) and $p < 0.05$ (**)) at the 17, and 18 tail somite stage.

Note that the delayed onset pattern was not observed for every gene that became male-enriched over this 24-hour period. For example, of the 208 probes that were male-enriched starting at E11.6 in 129S1, a few (*Gstm2*, *Etv5*, *Gas7*, and *Mybphl*) became sexually-dimorphic *earlier* in B6. Similarly, of the 75 genes that showed male-enrichment in B6 at E11.6, 11% (n = 8, including S*chip1, Lpl, Socs2*) become male-enriched later in 129S1 or not at all before E12.0. This indicates that although much of the male differentiation pathway is delayed in B6, this pattern is unlikely to be due to a more general delay in gonad differentiation.

Female-enriched genes exhibited a similar significant ~5 hour delay in B6. However, this strain delay stems from the later repression of female genes in XY gonads (Figure 24B, lower panel). As a consequence, B6 XY gonads are exposed to higher levels of female pathway genes for a longer period relative to 129S1 XY gonads. For example, of the 166 probes that become female-enriched in 129S1 starting at E11.6, only six exhibit a similar expression pattern in B6, while 70 become female-enriched ~5 hours later, another 43 probes become dimorphic at E12.0, and 47 fail to reach a sexually-dimorphic state by E12.0 (Figure 24C, lower panel). Similar to the male-enriched genes, 62.4% of female-enriched genes that become dimorphic at the same time in both strains show a higher fold difference in 129S1 than in B6 (Figure 25C, $p < 0.005$).

In summary, it is likely that the increased sensitivity of B6 XY gonads to sex reversal stems from a delay in the activation of male pathway genes downstream of *Sox9*, combined with a consequent delay in the repression of female pathway genes.

4.3.6 Using time course analysis to identify candidate genes in eQTL regions

In addition to providing a more comprehensive view of the global transcription dynamics driving male and female sex determination in the robust 129S1 and sensitive B6 strain backgrounds, this fine temporal expression data provided a means for narrowing eQTLs to identify novel regulators of sex determination. In previously published work, we mapped 19 regions of the genome in an F2 intercross population where genetic variation between B6 and 129S1 was correlated with differences in gene expression for one or more genes associated with sex determination (Munger et al. 2009). Eight of these regions were correlated with the expression of multiple genes, yet none of these prominent "trans-band eQTLs" harbored an obvious candidate gene with a known role in the sex determination process. Unfortunately, most of the eQTL regions identified in this initial coarse mapping were too large to functionally test every gene in the interval.

We established filtering criteria based on temporal strain expression and genomic data to prioritize candidate genes within the eight trans-band eQTLs (Table 8). Briefly, protein-coding genes in the interval were considered as candidates if they were expressed at one or more time points between E11.0-E12.0 in XY samples (eQTLs were mapped only in XY samples; therefore, the causative gene underlying an eQTL should be expressed in the XY gonad). Based on this list, we analyzed each candidate within the interval for strain differences in expression levels or time of onset, and prioritized genes with strain-dimorphic patterns. We investigated whether each gene harbored one or

more polymorphisms (SNPs, insertion/deletions) that differed between B6 and 129S1 and might affect its expression or function. Only those genes with characterized variation within 10kb up- and down-stream of the transcription start site (TSS) were prioritized for further analysis. Finally, we interrogated the Mammalian Phenotype (MP) browser to identify any genes in the region with a characterized knockout phenotype affecting sex determination (MP: 0002210, abnormal sex determination), or a known relationship with any of the target genes it was predicted to regulate. We tailored our candidate search strategy to each individual eQTL, and exploited prior information about the expression or function of the target genes for that region. Thus, we expected that genes involved in regulating early gonadogenesis genes would be expressed in both sexes at an early stage, whereas those regulating the male or female pathway would be more likely to exhibit sexually dimorphic gene expression.

Table 8: Identification of candidate genes in prominent trans-band eQTLs based on dynamic expression patterns

eQTL Chr: interval (Genes in MGI)	Controlled Transcripts	Expression pattern/ Function of transcripts	Expressed in XX/XY gonad	Sexually Dimorphic	Strain Dimorphic	Abnormal SD phenotype (MP:0002210)	Best Candidate(s)
1: 33-49cM (364)	*Sry, Sox9, Fgf9, Ptgds, Cbln1*	Male pathway	97	11	32	4	*Serpine2, Inha, Myl1, Igfbp5, Efhd1*
3: 65-76cM (60)	*Fog2 (Zfpm2), Ctnnb1, Pld1, SF1 (Nr5a1), Dapk1, Dock4, Cbln1, Rec8L1 (Rec8), Fgf9, Rspo1, Col9a3, Smpdl3b, Gata4, Socs2, Wt1, Asns*	Early gonadogenesis Female pathway Male pathway	20	6	13	0	***Lmo4***, *Gbp1/2/3, Ccbl2, Bcl10*
5: 26-46cM (276)	*Sphk1, Mmd2, Trim47, Dhh, Tpd52l1, Serpine2*	Male-enriched	61	12	15	5	*Pdgfra, Kit, Ppargc1a, Igfbp7*
12: 33-49cM (219)	*Rpgrip1, Wt1, Dock4, Fog2 (Zfpm2), Axin2, Pld1, Pdgfd, Ctnnb1, Dax1 (Nr0b1)*	Early gonadogenesis Female pathway	72	9	17	3	*Gtf2a1, Hspa2, Mlh3*
14: 31-39cM (174)	*Cbln1, Serpine2, Gng13*	Male-enriched	58	5	23	6	*Gata4, Ptk2b, Piwil2*
15: 36-58cM (526)	*SF1 (Nr5a1), Cst9, Pglyrp1, Rec8L1 (Rec8), Dtna, Mmd2, Centb1 (Acap1), Tpd52l1, Defb19*	Early gonadogenesis Female-enriched Male-enriched	208	29	77	13	*Sp1, Dhh, Amhr2, Sbf1, Smc1b, Mov10l1, Pfdn5, Pick1*
17: 19-39cM (367)	*Taf7l, Defb19, Ren1, Slco3a1, Smoc2, Pglyrp1*	Male-enriched Female-enriched	123	11	37	5	*Txndc2, Dazl, Rab31*
19: 36-41cM (108)	*Fst, Slco3a1, Smpdl3b*	Female-enriched Male-enriched	49	6	22	1	*Tmem180, Sema4g, Pax2*

Multiple criteria were established to identify the best candidate genes in the trans-band eQTLs mapped previously. First, the putative regulatory gene must be expressed above background at or before E11.5 to exert any effects on downstream target genes. Genes implicated in the list of target genes that exhibited a sexually-dimorphic expression pattern consistent with the sexual differentiation pathway (male or female) were prioritized. Genes that met both of the above criteria, and exhibited strain expression differences (either in overall levels of transcript abundance or in timing of onset of sexual dimorphism) in a pattern consistent with the observed allelic effects for that eQTL, were prioritized as the highest candidates. A few eQTLs harbor genes in

which abnormal sex determination phenotypes have been noted in the null mutant mouse, and these genes were given similar high priority.

In total, for the eight prominent trans-band eQTLs mapped, of which the average interval contains ~300 genes (range = 60-526 genes), we narrowed each down to, at most, eight promising candidates. Of particular interest, the distal Chr 3 region was strongly associated with the expression of nearly one-third of all the genes in our previous mapping study , including known regulators of early gonadogenesis (*Fog2*/*Zfpm2*, *SF1*/*Nr5a1*, *Gata4*, *Wt1*, and *Ctnnb1*) and both the female (*Ctnnb1*, *Rspo1*) and male (*Fgf9*) differentiation pathways. A total of six genes were identified as candidates based on their strain-dimorphic expression patterns, but only the transcription cofactor *Lmo4* (lim domain only 4) exhibited a dynamic pattern consistent with a role early in both pathways and an additional male-specific role downstream of the sexual fate decision. *Lmo4* is expressed at similarly high levels in both sexes until E11.4 (Figure 6A), and becomes male-enriched as early as E11.6. Importantly, while *Lmo4* was up-regulated in B6 XY gonads at the same time as in 129S1, there was a significant strain effect with expression in 129S1 being higher. This observation is consistent with the observed allelic effect for the eQTL (B6129SF2 gonads that were homozygous for the 129S1 allele exhibited higher expression of target genes). Finally, there is a significant amount of genetic variation in *Lmo4* between 129S1 and B6, including an insertion in the 3′ UTR in 129S1 as well as multiple intronic SNPs and indels (Keane et al. 2011). Based on these selection criteria, we elected to focus on developing a functional assay for *Lmo4*.

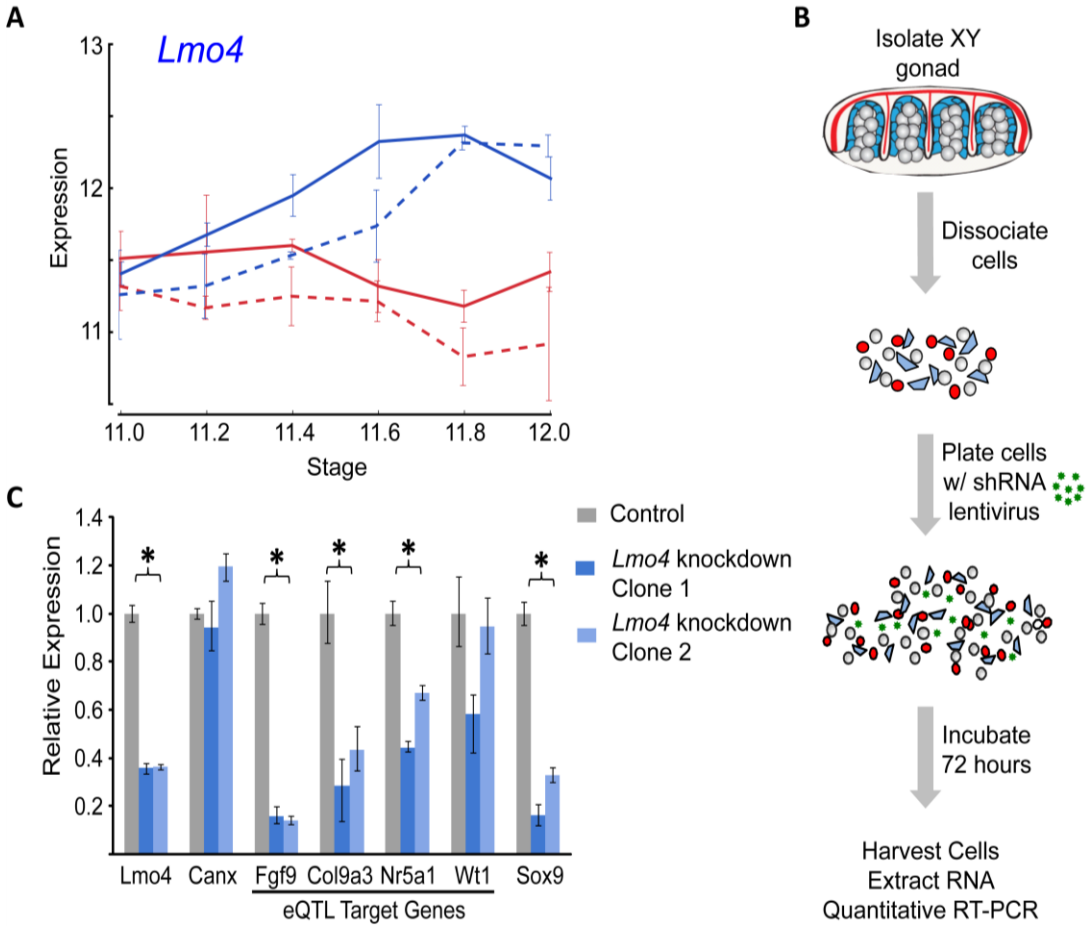


Figure 27: Validation of *Lmo4* as a novel regulator of gene expression in the fetal gonad.

(**A**) *Lmo4* exhibits an expression pattern indicative of a role in sex determination and consistent with expectations for a gene underlying a strain eQTL regulating early sex determination and male pathway genes. It is expressed at similar levels in XY and XX gonads before E11.6, becomes enriched in XY gonads as early as E11.6 in both strains, and shows reduced expression levels in B6 (dashed lines), consistent with the observed allelic effects of the Chromosome 3 eQTL. (**B**) E12.5 XY gonads were dissected free of the mesonephroi, pooled by sex, dissociated into single cell suspensions, plated

on tissue culture plates at t=0 with lentiviral particles containing shRNA targeted to the candidate gene of interest, and cultured for 72 hours. Quantitative RT-PCR was conducted to assay expression of predicted targets. (**C**) Lentiviral shRNA-mediated knockdown of Lmo4 in cultured XY primary gonadal cells resulted in the consistent down-regulation of multiple Chromosome 3 eQTL target genes relative to the non-targeting control (gray bar) using two different shRNA hairpins targeting Lmo4 (light/medium blue bars in graph). Expression was normalized to the housekeeping gene *Gapdh*. Both male pathway genes, *Fgf9* and *Col9a3*, were significantly down-regulated following Lmo4 knockdown with both clones. Similarly, one of the putative targets with a role in early gonadogenesis, *SF1/Nr5a1*, was significantly reduced, however expression of the other gene involved in early gonadogenesis, *Wt1*, was not significantly affected by *Lmo4* knockdown. The important male pathway regulator *Sox9* was found to be significantly down-regulated as a result of *Lmo4* knockdown. *Canx* (a second normalization gene not predicted to be a target of LMO4) showed no difference in expression compared to the control. Error bars show minimum and maximum expression. Significance was calculated by comparing control and data across all independent runs.

4.3.7 Validation of Lmo4 as the causative gene underlying the trans-band eQTL on distal Chromosome 3.

Historically, moving from a list of candidate genes to a validated quantitative trait gene (QTG) has represented the largest hurdle (in both resources and time) to success in complex trait mapping studies in the mouse. To address this problem, we optimized a lentivirus-mediated shRNA delivery method to artificially silence candidate regulatory genes with high efficiency in dissociated gonad primary cells from E12.5 XY gonads (Figure 27B). As a positive control, we utilized pre-designed and validated shRNA clones (Sigma MISSION) packaged in lentiviral vectors to silence *Sox9* expression in primary gonadal cell culture, and quantified the expression of known downstream targets (Figure 28). Lentiviral-mediated knockdown resulted in a nearly 80% reduction in *Sox9* expression relative to a non-targeting control sample. Two of the three known targets (direct or indirect) of SOX9, *Amh* (De Santa Barbara et al. 1998) and *Fgf9* (Kim et al. 2006), were down-regulated significantly following *Sox9* knockdown. *Ptgds*, the third known target of SOX9 (Wilhelm et al. 2007), is not expressed at high levels in cultured gonad primary cells, and we could not detect a change in *Ptgds* expression following *Sox9* knockdown. However, a marker of the female pathway, *Fst* (Yao et al. 2004), showed a significant and greater than 2-fold up-regulation in this assay ($p < 0.016$). Thus, this *in vitro* assay recapitulates well-characterized genetic interactions that occur in the gonad *in vivo.*

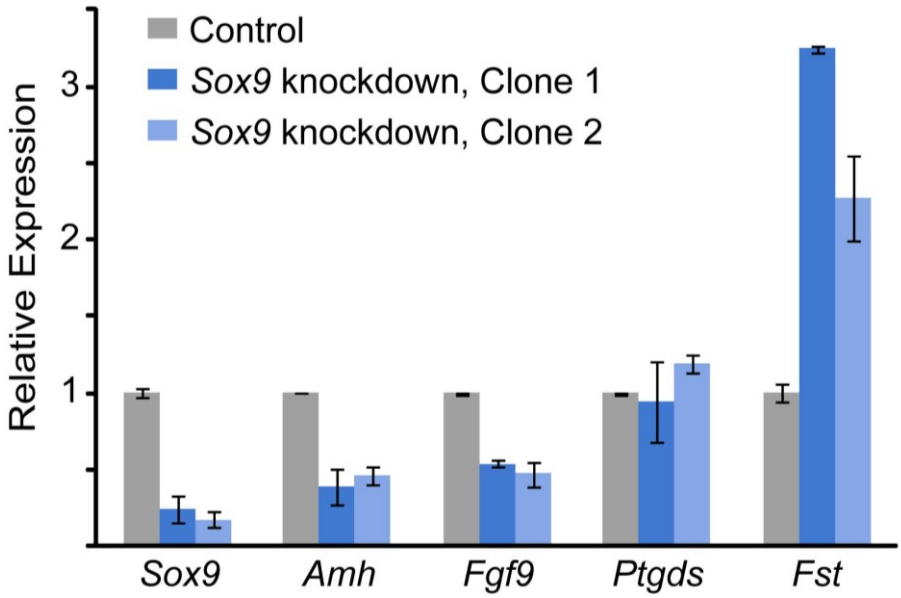


Figure 28: Lentiviral mediated knockdown of *Sox9* in gonad primary cell culture results in down-regulation of male-enriched genes.

Knockdown of *Sox9* resulted in down-regulation of known male-enriched genes such as *Amh* and *Fgf9* and up-regulation of female-enriched gene *Fst*. However, *Ptgds*, a known male-enriched gene (Wilhelm et al. 2007), does not show down-regulation.

We extended our analysis to test *Lmo4* as a candidate regulator underlying the Chr 3 eQTL. We silenced *Lmo4* expression to 36% of a non-targeted control ($p < 0.001$) (Figure 6C). Although the degree of knockdown was relatively modest, it was observed consistently in three independent trials and with two shRNA clones. Importantly, silencing *Lmo4* expression by two-thirds resulted in the consistent, significant down-regulation ($p<0.05$) of three of the four putative eQTL targets measured by qRT-PCR. *Fgf9*, *Col9a3*, and *SF1/Nr5a1* were significantly down-regulated by both shRNAs (Figure 6C). Down-regulation of a fourth target of the Chr 3 eQTL, *Wt1*, was not statistically significant ($p < 0.13$). Interestingly, although it was not identified as a Chr 3 eQTL target in our original mapping study, *Sox9* expression is significantly down-regulated following *Lmo4* knockdown by both shRNA clones ($p < 0.001$). Note that the predicted targets (*Fgf9*, *Col9a3*, and *SF1/Nr5a1*) of the Chr 3 eQTL region are those that are affected by the different alleles in B6 and 129S1. Even though *Sox9* did not map as a target of the Chr 3 eQTL in our study, it might not be differentially regulated by *Lmo4* between B6 and 129S1, yet could still be a target of *Lmo4.* In total, these experiments provide strong support for *Lmo4* as the transcriptional regulator underlying the Chr 3 trans-band eQTL and may reveal additional regulatory interactions that were undetected in the eQTL mapping study.

4.4 Discussion

4.4.1 Gonadal sex determination is orchestrated by a highly dynamic transcriptome

Previous microarray studies profiled transcript abundance in whole gonads or isolated cell populations at two or more time points before and after the sex determination decision (Nef et al. 2005; Small et al. 2005; Beverdam and Koopman 2006; Jameson et al. 2012b). These datasets served as important resources for the field. However, the temporal resolution around the critical stage of sex determination was limited in all but one study to 24-hour intervals (Nef sampled at E10.5, E11.0, and E11.5). It was evident from these earlier studies that the gonad transcriptome changed very little between E10.5-E11.0, that the difference between E11.0-E11.5 was significant, and between E11.5-E12.5 the testis and ovarian transcriptomes are highly sexually dimorphic. We predicted that information about the sequential order of gene activation/repression during the E11.0-E12.0 window would be valuable. Using data from the fine time course, we designed an HMM to precisely separate genes based on their position in the transcriptional cascade. As opposed to other clustering methods such as k-means clustering (Tamayo et al. 1999), HMMs are able to both account for the time dependence in the data and exploit this layer of information to identify patterns often obscured by noise prevalent in microarray data. We note that this HMM can be readily extended to time course expression analyses in other systems.

Our analysis identified waves of sexually-dimorphic gene expression in the 24-hour window following the onset of *Sry* expression, which suggest regulatory cascades. We note that while the HMM identifies dimorphically expressed genes as being dimorphically expressed at distinct time points, this is a result of the sampling times of our transcriptome analysis. Finer sampling in this window is likely to reveal that genes grouped together at a time point show minor differences in timing of the onset of dimorphism. Previous work indicated that the supporting cell lineage is the first lineage in the gonad to show sexually dimorphic expression followed by other gonadal cell lineages after E11.5 (Jameson et al. 2012b). Consistent with this, over half (58%) of the genes we identified that became sexually dimorphic prior to E11.8 could be specifically assigned to the supporting cell lineage prior to E11.8 with 5% showing dimorphism in germ cells. The discrepancies in the overlap are likely due to the increased sensitivity of the HMM to identify dimorphically expressed genes and the conservative measure of dimorphic expression used in the cell-type specific expression study.

As expected, we observed a strong signature of gene activation associated with up-regulation of the testis pathway in XY gonads. However, we were surprised by the extent of the repressive program that silences female pathway-associated genes in the XY gonad following activation of the male pathway. Testing of candidates from our study will be an important step towards identifying the factors responsible for these

patterns of expression in the male and female program. Based on their early onset of sexually-dimorphic expression, several genes are promising candidates to play an early regulatory role in the male and female pathways. *Sox13* is a member of the SOX protein family that lacks an activation domain but can repress Wnt signaling by forming a complex with the β-catenin cofactor, TCF1 (Melichar et al. 2007; Marfil et al. 2010). *Mef2c* is activated in the XY gonad at E11.6 and has been shown to interact with *Sox9* in chondrocytes (Dy et al. 2012). Among genes that showed female-enrichment, *Gtf2a1,* a general transcription factor that is part of the initiation complex for PolII recruitment (DeJong et al. 1995), became dimorphic at E11.4 in the XX supporting cell lineage and *Tcea3*, a known PolII elongation factor became dimorphic at E11.6. Interestingly, female-enriched TFs such as *Zfp277, Runx1, Lef1, Lhx9* and *Msx1* were strongly down-regulated in XY gonads. Conversely, *Irx3* showed strong activation in XX gonads, and has been predicted to have a function during ovarian differentiation independent of *Foxl2* and *Wnt4* (Garcia-Ortiz et al. 2009).

4.4.2 Sensitivity to sex reversal in B6 stems from the delayed onset of the male pathway downstream of Sox9.

Strain differences in resistance to sex reversal upon perturbation of the sex determination network have been the focus of several studies. The importance of the timing of the antagonistic testis and ovarian programs to B6-associated sex reversal was

first proposed by Eicher in 1983 (Eicher and Washburn 1983), based on the finding that introduction of a *Mus domesticus* Y chromosome (Y^{Dom}, or Y^{POS}) onto a B6 genetic background led to sex reversal (Eicher et al. 1982). Sex reversal in this case was later shown to be associated with the delayed onset of *Sry* (Bullejos and Koopman 2005). However, this work did not explain why the B6 strain is more susceptible to sex reversal in cases where a weak allele of *Sry* is not involved (Eicher et al. 1996).

Here we showed that the onset of sexually dimorphic gene expression was delayed by approximately five hours in the "unperturbed" (i.e. wildtype) B6 strain compared to 129S1. This delay is consistent across the cascade starting from E11.4 with genes that are both up- and down-regulated in XY gonads. Interestingly, we detected no significant difference by qRT-PCR in the level of *Sry* expression between the strains; however we cannot rule out a difference in the onset of *Sry* expression prior to the window of our analysis. Nonetheless, *Sox9* is up-regulated at the same stage (E11.2) in B6 and 129S1. Despite this agreement, the activation of many downstream genes in the male pathway is delayed in B6. In our previous microarray comparison of B6 and 129S1 testes at E11.5 (Munger et al. 2009), *Sox9* was found to be enriched in B6 relative to 129S1 XY gonads at E11.5, in contrast to the current study, where *Sox9* levels are lower in B6 until E11.8-E12.0 (Figure S7B). This discrepancy may stem from small developmental staging differences in the pooled gonads used in the previous study, as *Sox9* levels are

changing very rapidly at E11.5. However, as this and other recent studies (Hiramatsu et al. 2009; Correa et al. 2012) illustrate, the system may be sensitive to minor fluctuations in gene expression between E11.0-E11.5. Thus, the slightly lower level of *Sox9* expression that we detected in B6 relative to 129S1 XY gonads might contribute to the delayed onset timing of downstream genes in B6. In addition, our fine time course data here helps explain the previously observed higher expression of female-enriched genes in B6 compared to 129S1. Specifically, the observed difference at E11.5 is a consequence of the delayed repression of female-enriched genes in B6 XY gonads.

4.4.3 Prediction and validation of Lmo4 as a novel regulator of sex determination.

Previous transcriptome and genetic mapping studies produced gene lists or large intervals with candidate regulators of sex determination (Eicher et al. 1996; Nef et al. 2005; Small et al. 2005; Beverdam and Koopman 2006; Bouma et al. 2007; Nikolova et al. 2008; Munger et al. 2009; Bouma et al. 2010). The bottleneck in applying the results of these studies has been the inability to prioritize between several dozen candidates and then test these candidates in a manner that is inexpensive and efficient. We have addressed both these deficiencies in the current study. We used our fine time course dataset in conjunction with a previous eQTL study and cell-type expression data to identify candidate regulators of sex determination.

To overcome the hurdle of testing candidate regulators, we developed an RNAi assay to silence the expression of candidate genes, and then monitored the expression of putative downstream target genes after knockdown. As predicted, shRNA-mediated silencing of the transcription cofactor *Lmo4* resulted in the down-regulation of known important regulators of early gonadogenesis (*SF1/Nr5a1*) and the male pathway (*Sox9* and *Fgf9*). This provides strong evidence that in addition to previously characterized roles during development in the neural tube (Hahm et al. 2004; Tse et al. 2004; Lee et al. 2005), neural crest (Ochoa et al. 2012), cortex (Asprer et al. 2011), and thymus (Michell et al. 2010), *Lmo4* is also a regulator of sex determination in the gonad. However, this does not preclude the possibility that other genes on distal Chr 3 have roles during sex determination and control the expression of one or more of the 16 eQTL target genes. To point, four of the other candidate regulators identified in this region (*Gbp1/2/3* and *Ccbl2*) are expressed at similar high levels in both sexes before E11.4, and then become down-regulated specifically in XY gonads at or after E11.6. This pattern predicts a role for these genes in the female differentiation pathway. Future assays to over-express these candidates in XY primary cells or silence them in XX primary cells will assess their potential as regulators for one or more of the Chr 3 eQTL target genes.

5. Setting up ChIP-seq in the gonad

5.1 Summary

While being able to conduct ChIP-seq in the gonad would be a crucial step in deepening our understanding of the transcriptional regulation that is carried out during the fate commitment process, it has been a significantly challenging task. We took a methodical approach to solving the problems involved. We started by optimizing the sonication settings to produce DNA fragments in the size range between 200-1000bp. We then used a mouse myoblast cell line (C2C12) for which ChIP-seq datasets had been published, to evaluate the various parameters in the assay, with a focus on using small numbers of cells (< 500,000 cells/IP). Finally, we successfully prepared libraries from the amounts of DNA we expected to obtain from our ChIP-seq experiments in gonadal cells. These steps were crucial in preparing us for conducting ChIP-seq in the supporting cells.

5.2 Introduction

ChIP is a particularly useful technique that reveals where a protein is localized on the genome (Furey 2012). This technique has been used across a wide range of biological systems to reveal the specific patterns of a variety of DNA binding proteins, especially TFs. Given the vast scale of dynamic transcriptional regulation that we observe in the gonad as it differentiates to a testis or an ovary, it is extremely important that we can understand which regions of the genome drive these patterns of expression. However, setting up ChIP in the gonad has been problematic.

The major steps in ChIP are as follows. First, after samples are collected, proteins are cross-linked to DNA, typically with formaldehyde. In some cases, where the proteins are tightly bound to the DNA, such as histones, cross-linking has been shown to be not strictly necessary (Cuddapah et al. 2009). Second, the cross-linked protein-DNA complexes have to be sheared into small fragments between 200-1000bp. This is usually accomplished by sonicating the sample in a lysis buffer or treating with the right concentration of MNase to shear the chromatin. Note that the smaller the fragment length, the higher resolution one has of where the protein is bound. However, the stress of excessive sonication or digestion of the sample might lead to the protein not remaining bound to the DNA. Therefore, a fine balance between resolution and the ability to immunoprecipitate sufficient DNA-bound protein needs to be achieved. Third,

the sheared chromatin is incubated with an antibody specific to the protein of interest. This could be TFs, histone variants, or proteins with specific post-translational modifications, such as histone marks. After sufficient time has been allowed for the antibody to bind to the protein, secondary antibodies, that are themselves bound to magnetic or agarose beads, are used to capture the primary antibody bound protein DNA complexes. Fourth, a series of washes are applied to remove non-specific binding. Fifth, the cross-links between the protein and DNA are reversed and the remaining protein and RNA are digested with proteinases and RNases, respectively. Finally, the DNA is purified using either a phenol-chloroform extraction or a purification column.

Concurrent to this process different controls are conducted. One control is called the input or whole cell extract control. This is collected by simply purifying DNA from sheared chromatin without incubation with any antibodies. An alternative control that provides a measure of the non-specific binding of the antibodies is the IgG control. The IgG antibody is not expected to bind any of the chromatin bound protein and is used in exactly the same fashion as the antibody to the protein of interest.

The steps in this process have to be optimized carefully for different types of cells and different antibodies. Specifically, sonication cycles and lysis buffers might have to be optimized for different cell types and numbers. Additionally, the antibody amount

and the number of washes are important variables that can be tweaked to achieve better signal-to-noise ratio in experiments.

In aiming to set up ChIP-seq in the supporting cells of the gonad, the main problem is obtaining sufficient biological material. While typical ChIP assays use ~5M cells/IP, FAC-sorts for supporting cells with *Sox9-CFP* positive male gonads at E13.5 can roughly provide ~15,000 cells per embryo. While this already implies that >300 pairs of testis have to be collected, the situation is far worse with E13.5 ovaries, with only ~1,500 supporting cells being FAC-sorted per embryo. Therefore, one requires >3,000 XX embryo to be dissected to get enough material to obtain the necessary number of XX supporting cells! The situation is even tougher with cells of the bipotential gonad, as the dissections of gonads prior to E11.0 are significantly tougher.

Early attempts at ChIP by a postdoctoral associate, Danielle Maatouk, in the Capel lab were focused on identifying the binding sites of beta-catenin, a co-factor that is downstream of Wnt signaling. Since, it was observed that Wnt signaling is important in ovarian development and that stabilization of beta catenin was at least one of the main mechanisms for this effect, understanding the binding locations of the co-factor is an important step towards elucidating the role of *Wnt4* in ovarian development (Maatouk et al. 2008). Using whole female gonads from several embryos, the first step was to sonicate the samples to obtain a smear of DNA in the 200-1000bp range. However,

despite trying various sonication protocols, including using Bioruptors and probe sonicators, the size of the DNA fragments was seen to be consistently above the required range of fragments for ChIP experiments. This immediately created hurdles for obtaining useful ChIP data and this approach did not bear much fruit. With the benefit of hindsight, the reason for the lack of DNA fragments in the required fragment range might have been that whole tissue was being used. As a result, the sonication has to overcome the extra-cellular matrix of the gonad tissue. Further, the mix of cell types in the gonad, each with a different resistance to sonication might have lead to observing smears over much larger size ranges.

An attempt by Danielle Maatouk that proved far more effective was using a carrier ChIP protocol for low cell numbers (O'Neill et al. 2006; Rugg-Gunn et al. 2010; Zwart et al. 2013). This protocol differs from the standard ChIP protocol in two ways. First, MNase digestion is used to shear chromatin in cells that have not been cross-linked. Second, yeast or Drosophila cells are spiked in to the biological material of interest. Following this, species specific primers are designed for the sites of interest to identify enrichment by qPCR. The rationale behind adding carrier chromatin is to reduce the background observed when one only has a few cells (<1000 cells) to perform ChIP. Alternatively, recent protocols have attempted to address the problem of higher

background with low sample quantity by optimizing the amount of antibody for specific cell numbers (Adli et al. 2010; Adli and Bernstein 2011).

We set ourselves the goal of developing ChIP-seq to use with supporting cells from the gonad. This not only would allow us to answer questions of interest such as which regions are active enhancers in supporting cells, but also open a whole host of questions that are well suited to be answered by the application of ChIP. In establishing this technique, I relied heavily on several recent protocols that were developed with the goal of collecting ChIP-seq data from small cell numbers (Dahl and Collas 2009; Adli and Bernstein 2011; Sachs et al. 2013).

5.3 Results

5.3.1 Sonication efficiency

The first step was to conduct tests to ascertain what combination of lysis buffer and sonication settings was to be used to shear chromatin to get DNA size fragments in the required size range. There are several settings that one can tweak for sonication. These include the output power of the sonicator, the duty cycle (the percent on time/second of the pulse), the number of pulses per sonication, the wait time after each round of sonication and the number of cycles of sonication to use.

Since there were a high number of parameters to optimize, it seemed unwise to begin with the supporting cells of the gonad, as collecting enough cells would be extremely time inefficient. Therefore, we used HEK293T cells which could be easily cultured, to test out these parameters. The goal with these experiments was to sonicate equal numbers of cells (~1M cells) with different settings, reverse the cross-links, purify DNA and run gels to evaluate the different size fragments. After attempting several different sonication protocols, it was observed that on a Branson sonicator, for HEK293T cells, between 8 and 16 cycles of 30s of sonication, followed by 60s of rest time provided the required fragmentation. Each sonication was performed at output power 3 with a duty cycle of 30%.

Varying these settings, several useful observations were made. First, it was important to keep the tube on ice while sonicating. In addition, 70% ethanol was added to the ice to keep the sonication tube below freezing. Second, using a clamp to hold the tube during sonication was important to obtain consistent results. Attempts at manually holding the tube at exactly the same height for extended periods of time proved extremely difficult. Further, minor movements resulted in the tip probe touching the sides of the tube or foaming of the lysate. The use of a clamp allowed sonication to be conducted at exactly the same height and position without the need for interference from the experimenter after setting up the apparatus. Third, extended periods of continuous sonication resulted in foaming of the lysate. This likely was due to over-heating of the sample.

We used 16 cycles of sonication at the above settings with C2C12 cells (used for optimization of the assay) and discovered that we could achieve fragments of the right size as analyzed by an Agilent tapestation (Figure 29). Eventually, the same setting was tested first on E17.5 XY supporting cells and then used on the E13.5 XY supporting cells.

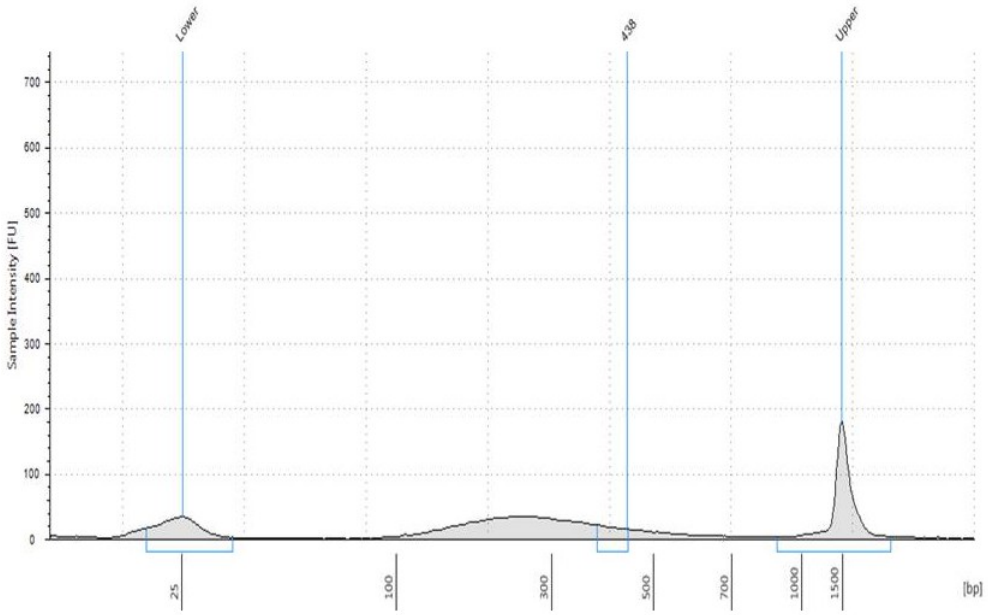


Figure 29: Fragment size from sonicated C2C12 cells

100,000 C2C12 cells were sonicated in lysis buffer after cross-linking. Cross-links were reversed and DNA was purified. DNA was run on Agilent tapestation.

5.3.2 Picking the right cell type for testing

Ideally, we would have liked to evaluate the various ChIP protocols on the supporting cells of the gonad. However, since this would require significant collection of material, it would have been extremely time-inefficient. As a result, we attempted to identify a cell type that could be used to evaluate the various protocols.

Several criteria were used to select the cell type for testing ChIP. First, we wanted to test our protocols and antibodies against mouse cell lines, as our final samples would be from the mouse. This would eliminate any cross species variation in the binding of the antibody. Second, we needed a cell type in which we understood the 'ground truth' situation. This meant access to published ChIP-seq datasets for a number of proteins and post-translational modifications, so we could pick regions we knew were going to be either enriched or not for the proteins of our interest. With these criteria, we had two main candidates – ES cells and C2C12 cells. In choosing between these two cell types, we applied our third criteria – the ease and cost of culturing these cells. Since ES cells need significant effort to culture, and C2C12 cells are relatively easy, we decided to use C2C12 as the cells on which we would conduct the tests of the different antibodies and protocols. Crucially, genome-wide profiling data existed for this cell type for several enhancer correlated marks and proteins including, P300, RNA PolII, H3K4me1 and H3K27ac (Asp et al. 2011; Blum et al. 2012).

5.3.3 Attempts to use P300 to identify enhancers

P300 is a ubiquitously expressed histone acetyl transferase that, in addition to catalyzing the acetylation of several histone residues including H3K27ac, can acetylate residues on several non-histone proteins (Ogryzko et al. 1996; Struhl 1998). While the specific mechanism by which this protein has an effect on transcriptional regulation is still unclear, it likely involves a combination of being able to facilitate the communication between the distal regulatory elements and the promoter and histone acetylation functions.

Despite the lack of understanding regarding the precise mechanism of P300's action, the localization of this co-activator has been used in multiple contexts to discern cell-type specific and tissue specific enhancers (Visel et al. 2009; Blow et al. 2010). In one excellent validation study, the authors collected ChIP-seq data for P300 from limbs, forebrain and midbrain of mice at E11.5 (Visel et al. 2009). They then tested 86 different enhancer regions by making lacZ reporter mice. Of these, 87% of embryos were positive for enhancer activity specifically in the tissue type where P300 was found to bind the region. This is a far higher rate of success in identifying enhancers than has been achieved with other methods (Kantorovitz et al. 2009). Additionally, the peaks for P300 tend to be on the order of a few hundred base pairs which agrees well with size of known enhancers. This is in contrast to histone modification marks that are broad

kilobase sized regions surrounding the enhancer (Buecker and Wysocka 2012). For these reasons, we wished to conduct ChIP-seq with P300 on the supporting cells of the mouse gonad.

Our attempts with P300 ChIP on C2C12 cells were unsuccessful (Figure 30). Specifically, we noticed that we were unable to find enrichment for regions that were known to be positive from ChIP-seq datasets (Blum et al. 2012). In addition, we observed that the enrichment was not significantly greater than what could be observed for immunoprecipitation with the IgG control. Increasing the sample amount and the antibody concentration showed no improvement.

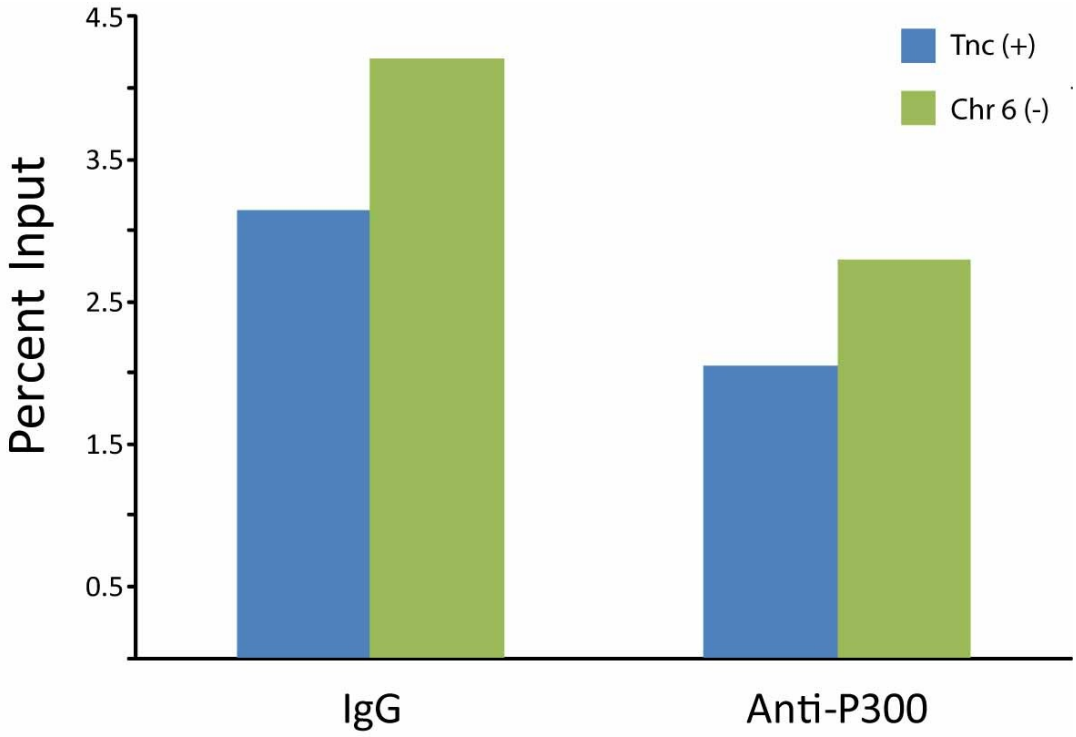


Figure 30: ChIP for P300 in 5M C2C12 cells

One possible reason for the failure of this approach could have been that P300 is a cofactor that is not tightly bound to the DNA directly. As a result, cross-linking of this protein to the DNA was likely highly inefficient. To overcome this, we attempted using a double cross-linking method, where proteins are first cross-linked, and this is followed by formaldehyde cross-linking to create protein-DNA covalent bonds (Zeng et al. 2006). Unfortunately, this method too, showed no improvement in enrichment of P300 binding for positive regions over the background.

After consulting with multiple researchers who have carried out P300 ChIP successfully, we learnt that P300 antibody shows considerable variability between batches and does not immunoprecipitate large quantities of DNA. This indicated that conducting ChIP-seq with small sample quantities would be significantly problematic as the amount of DNA retrieved might be too little to build libraries. Therefore, we decided to abandon trying to ChIP P300 and attempted to ChIP histone marks correlated with active enhancers.

5.3.4 ChIP for histone marks

Several histone marks have been shown to be correlated with enhancers. Initial work identified H3K4me1 as a mark that was enriched particularly in distal enhancers as opposed to the promoters of genes (Heintzman et al. 2007). More recently, it was

discovered that H3K4me1 marks two classes of enhancers (Creyghton et al. 2010; Rada-Iglesias et al. 2011). One class is called 'active' and these enhancers are expected to be participating in driving the expression of downstream genes in the specific cell type. These enhancers are seen to have the H3K27ac mark around these regions, likely added by P300 or other histone acetyl transferases. The other class of enhancers is called 'poised', and these enhancers are expected to allow for enhancer activity, but are currently not participating in driving the expression of downstream genes. These enhancers are likely important in the differentiation of the cell to other cell types, or to provide the competence to react to extracellular signaling or other environmental stimuli. These enhancers do not exhibit the H3K27ac mark.

We were able to successfully perform ChIP-qPCR with these two histone marks, H3K4me1 and H3K27ac, in C2C12 cells (Figure 31). We analyzed published ChIP-seq data to pick regions of the genome that should be enriched for these two marks and were successful in finding these regions enriched in our ChIP-qPCR experiments (Blum et al. 2012). We conducted several iterations of our ChIP experiment, to optimize the number of washes and the antibody amounts and were eventually able to produce reliable results with 400,000 cells/IP, which is an order of magnitude less than the number of cells typically used in ChIP assays.

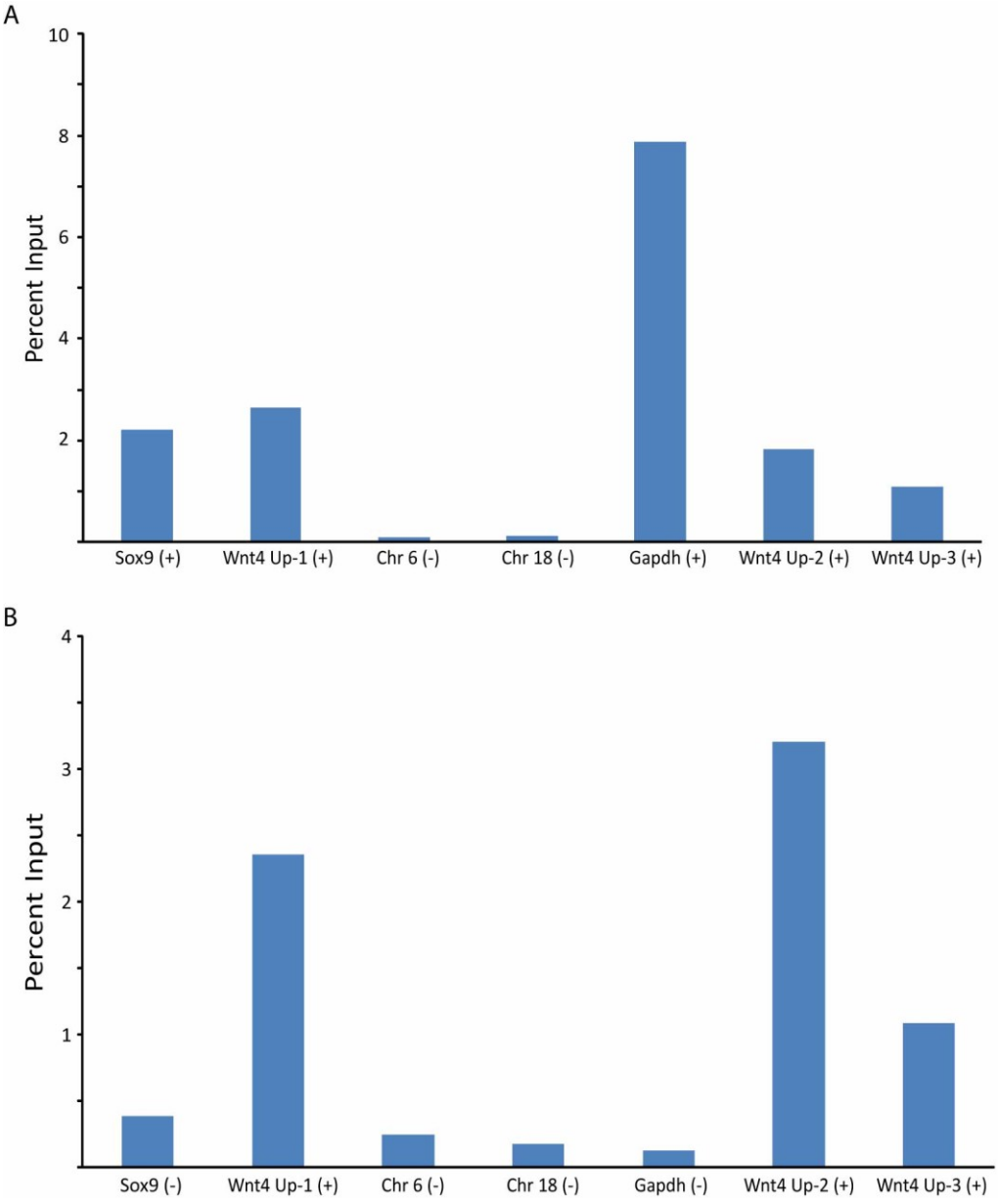


Figure 31: ChIP for H3K27ac and H3K4me1 in ~400K C2C12 cells

ChIP for H3K27ac (**A**) and H3K4me1 (**B**) with ~400,000 cells/IP was seen to work as expected. The different regions used to test the assay are shown in the X-axis. The sign in the parentheses indicate if the region was expected to be enriched (+) or not (-).

5.3.5 Library preparation and size selection

A significant hurdle to conducting ChIP-seq with small samples is that the amount of DNA in the IP is much lesser than the amount required for building a library that can be sequenced. To overcome this, researchers typically perform a whole genome amplification of the IP fragments and then use this amplified sample in a library preparation (Adli and Bernstein 2011; Ng et al. 2013). More recently, Rubicon's ThruPLEX kit has been used with extremely low amounts of starting material (Sachs et al. 2013). We contacted researchers who used this kit and were informed that it was easy to use without significant need for optimization. In addition, we analyzed publicly available data that had been published by using this kit for library preparation from small cell numbers and observed that the data were of a high quality. For these reasons, we chose the Rubicon ThruPLEX kit to conduct library preparations.

We used the input and ChIP samples from C2C12 cells to prepare libraries and found that the amount and the size distribution were close to the right size range (Figure 32). Upon consulting with the sequencing core, we were informed that ideally one would prefer a lesser enrichment of the larger size fragments to maximize the number of reads obtained from each lane. To do so, we used a SPRI bead size selection, where the concentration of the SPRI beads used determines the different size fragments that are selected for (DeAngelis et al. 1995).

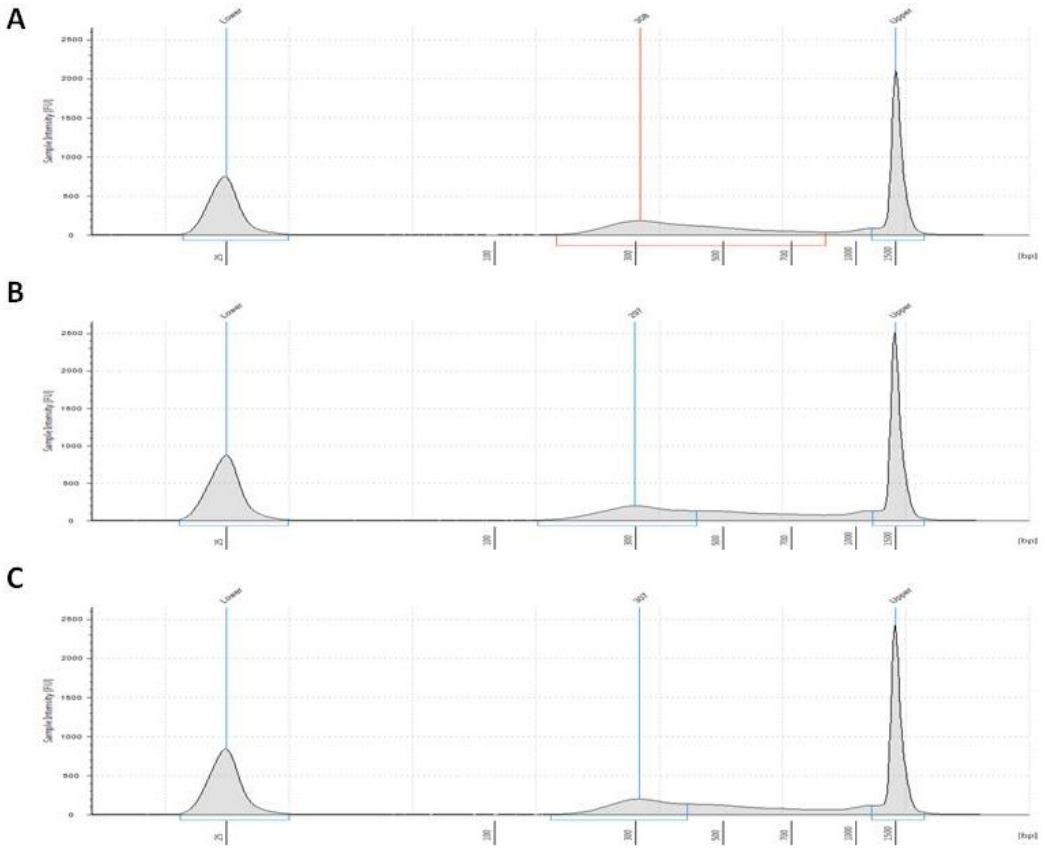


Figure 32: Library Preparation with Rubicon ThruPLEX FD kit for C2C12 cells

Library preparation for C2C12 cells from input (**A**), K4me1 (**B**) and K27ac (**C**) show a size distribution with most fragments between 150-800bp.

To optimize the concentration of SPRI beads used, we used a diluted DNA ladder. We added SPRI beads of different concentrations, which would bind to the large size fragments. The supernatant containing the low size fragments was cleaned using a high concentration of SPRI beads. The resulting low size DNA fragments were run on an Agilent tapestation and 0.6x was the concentration that was determined to be ideal in depleting high fragment sizes while keeping a large fraction of low size fragments. This protocol was applied the C2C12 trial library preparation and we observed libraries with better size distribution.

5.4 Discussion

The ability to conduct ChIP-seq is a particularly important step towards understanding the mechanistic underpinnings of transcriptional regulation in gonadal cells. Not only does the development of this assay open doors for identifying enhancers, one can use this technique to study several other problems. For example, this can be used to localize TFs across the genome as the gonad differentiates. This approach can also be used to identify the binding sites of protein complexes such as Polycomb to identify genes that are repressed.

In addition, to the typical challenges involved in conducting ChIP-seq, we had to overcome the limitations of the small amount of biological material we were able to obtain from embryonic gonads. As a result, we took a methodical approach to developing a ChIP assay for the gonad, by attempting several protocols with the C2C12 cells. Of importance was attempting to use small amounts of C2C12 cells and produce reliable and repeatable ChIP-qPCR data. In addition, the ability to prepare libraries from small amounts of DNA was a crucial step towards being able to successfully conduct ChIP-seq in gonadal cells.

6. Identifying active enhancers in XY supporting cells

6.1 Introduction

We wished to use the techniques we had developed with human cell lines to discover the TFs that play a regulatory role in Sertoli cell differentiation (Natarajan et al. 2012). Our approach in that study was to use expression data to first identify the cell-type specifically expressed genes. We have conducted two separate expression studies and have a detailed characterization of the genes that are differentially expressed in each of the cell types of the gonad (Jameson et al. 2012b; Munger et al. 2013).

We describe here a pilot study with E15.5 XY supporting cells. Specifically, we found that the DNase-seq assay successfully identified several putative regulatory regions. These regions show similar properties to those observed in human cell line DNase-seq data, indicating that our assay was successful. Further, we observed that DHS unique to Sertoli cells were enriched particularly around Sertoli cell specific expressed genes. Finally, we identified a previously unknown enhancer of Wt1 by using a transient transgenic embryo with a construct expressing LacZ under the control of a DHS that we identified in our study. Following this, we collected DNase-seq and ChIP-seq data for histone marks in E13.5 XY supporting cells to identify active enhancers. In addition to identifying the TFs involved in gene regulation, mutations in these distal

regulatory regions might also explain the as yet unexplained cases of human sex reversal.

6.2 Results

6.2.1 Conducting DNase-seq with E15.5 XY supporting cells

One of the main impediments to conducting DNase-seq in primary gonadal cells was insufficient material. To establish the feasibility of this approach, a previous postdoctoral associate in the lab, Danielle Maatouk, in collaboration with Greg Crawford, conducted a pilot project with E15.5 XY supporting cells. At this stage, one is able to collect more cells than at the earlier E13.5 stage and therefore this expedited the process of testing the feasibility of the assay. Specifically, they discovered that they could reduce the number of cells required from >10 M cells, to ~5M cells, a significant reduction in the number of cells. Crucially, this experiment was successful and ~80,000 DHSs were identified in the E15.5 supporting cells. To validate the veracity of the data we had collected, we characterized a few metrics.

We first looked to see if the DNase-seq peaks were found in the similar proportions in the different regions of the genome and at the right sizes (Figure 33 A, B). Crucially, as with the human data, we observed that a majority of DHS peaks (67%) lay in intergenic and intronic regions of the genome. There were ~18,000 peaks that overlapped promoters of ~13,000 genes. Further, as seen with the human data (Figure 3), the peaks overlapping promoter regions where significantly larger than those observed at other regions of the genome. In addition, Danielle Maatouk collected RNA-seq data

from E15.5 XY supporting cells and we observed a weak correlation between expression values and DNase-seq signal at the peak (Figure 33 C, D). Crucially, we observed a large variation of expression in genes with peaks at their promoters. This agrees well with our previous observation that while the lack of a promoter peak typically implies lack of expression, the presence of a peak can lead to a wide variety of expression outcomes.

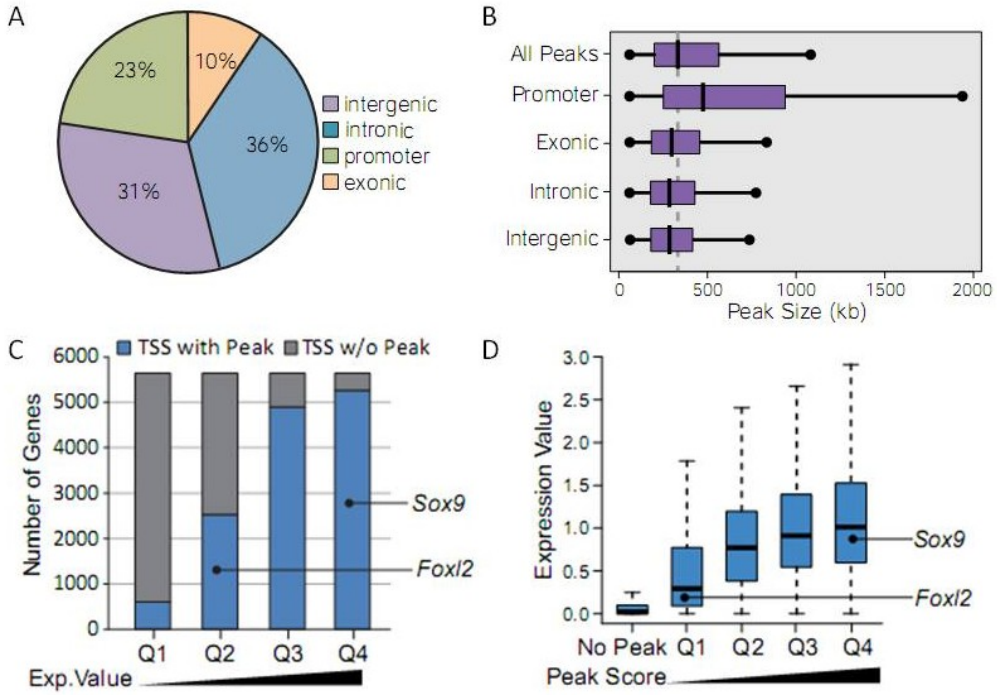


Figure 33: Validation of DNase-seq assay in E15.5 XY supporting cells

(**A**) Genomic locations of the DHS peaks. (**B**) Box plot showing the different sizes of peaks by their genomic locations. Promoter overlapping DHS peaks are seen to be significantly bigger than other peaks. (**C**) Expression values from RNA-seq were divided into quartiles and the number of genes with and without peaks at their promoters is shown. (**D**) Promoter DHS peaks were divided into quartiles based on the signal strength and expression values of the genes in these quartiles as box plots.

6.2.2 Unique DHSs are enriched near Sertoli cell specific expressed genes

In order to identify the DHSs that are crucial to drive expression in Sertoli cells, we divided the DHS peaks from E15.5 XY supporting cells as follows. By comparing our DHS data with data from 6 other cell lines, we asked in how many cell types each DHS was open in. Those DHSs that were open in all cell lines were termed common and those that were open just the Sertoli cells were termed unique DHS. Those DHSs that were open in a subset of cells were termed specific DHS. Figure 34A shows a good example of common and unique DHSs. We see that the promoter for *Sox9* is open in all cell types making it a common DHS. In contrast, the region upstream labeled TESCO, one of the few well characterized enhancers in Sertoli cells, is unique to Sertoli cells. This is in good agreement with previous work showing that cell-type specific DHSs are enriched for enhancers.

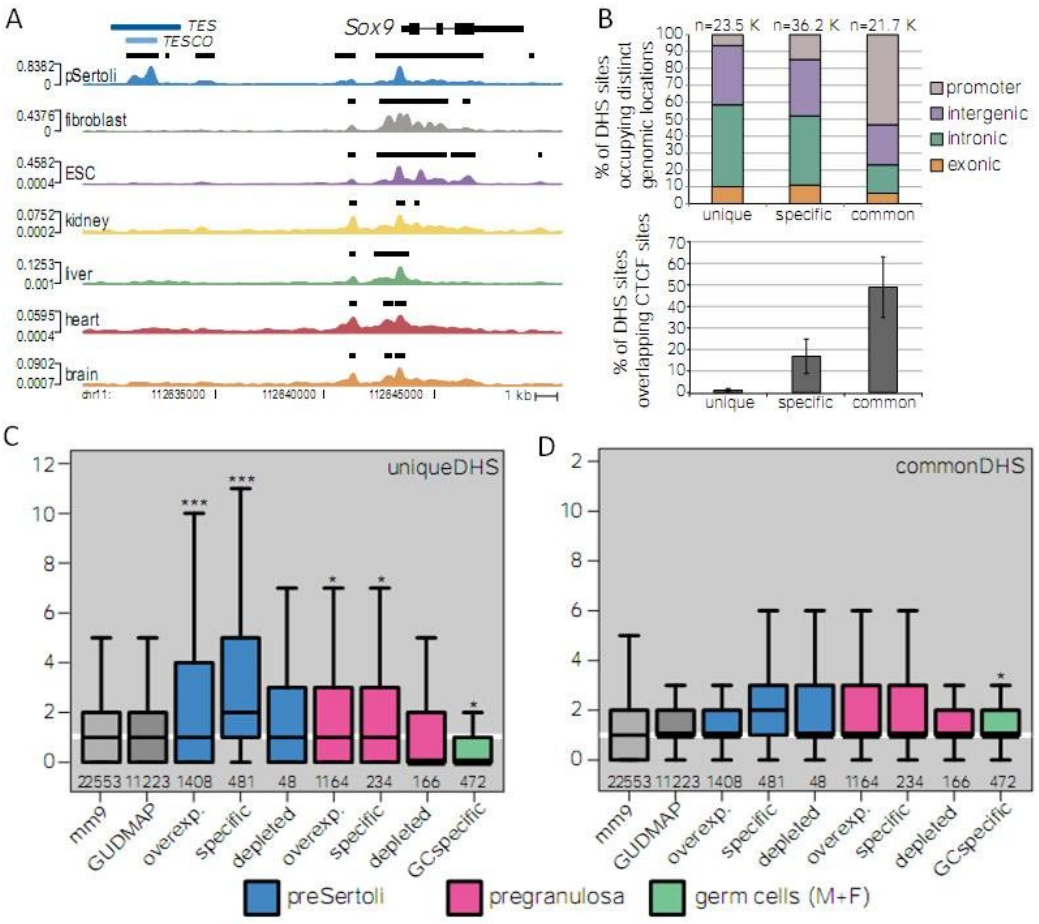


Figure 34: Sertoli cell unique DHSs are proximal to the Sertoli cell expressed genes

(**A**) Tracks showing DHSs from different cell types and tissues. The TES and TESCO elements were characterized previously (Sekido and Lovell-Badge 2008). This region is seen to be unique to Sertoli cells. (**B**) Different genomic locations of DHS based on the number of cell types they are open in. Unique DHSs are depleted for promoter DHSs, while common DHSs are enriched for promoter peaks (top panel) and CTCF peaks (bottom panel). (**C, D**) Proximity analysis for genes found enriched and depleted

in different cell types in the gonad from a previous study (Jameson et al. 2012b). (**C**) Unique DHSs are found enriched near the Sertoli cell specifically expressed genes and also moderately near Granulosa expressed genes compared to all genes. (**D**) Common DHSs do not show significant variation across genes expressed in different cell types. A two sided Mann-Whitney test was used to calculate significance. $p < 1e\text{-}3$ (*), $p < 1e\text{-}15$ (***).

We then analyzed the genomic locations of these three different groups of DHSs (Figure 34B). Here, we saw a striking difference between the three groups of DHSs with the majority of common DHSs being promoter peaks and >70% of them being found in genic regions. Interestingly, unique DHSs showed an opposite pattern with <10% being present in promoter DHSs. However, > 80% of these peaks are found in regions that are typically enriched for cell-type specific enhancer activity such as intergenic and intronic regions. We also found that 50% of common DHSs overlap a CTCF site found in ES cells. Further, unique DHSs are depleted for these sites.

A previous study in the Capel lab had investigated the cell-type specific expression profiles in different cell types in the gonad at the E13.5 stage (Jameson et al. 2012b). We enquired whether the chromatin landscape around the Sertoli cell specific genes was different from those around genes expressed in other cell types in the gonad. We found that while there were a significantly larger number of unique DHS peaks around these genes than that found at all genes identified as expressed, we saw no such difference for common DHSs (Figure 34 C, D). We further saw that germ cell enriched genes showed depletion of unique DHSs and a mild depletion in common DHSs.

Finally, we wished to confirm that at least a subset of these unique DHSs were enhancer regions and tested whether they drove enhancer expression. To do so, we picked a unique DHS region upstream of the *Wt1* promoter. We cloned this region

upstream of an *Hsp68-LacZ* gene. We made transient transgenic mouse with this construct and observed that this region drove expression in Sertoli cells in 6/7 embryos. Further, we saw expression in the somatic cells of the ovary in 11/12 embryos, which agrees with the expression of *Wt1* in both sexes in the early stages. In conclusion, this region represents a bona fide enhancer of the *Wt1* gene in Sertoli cells.

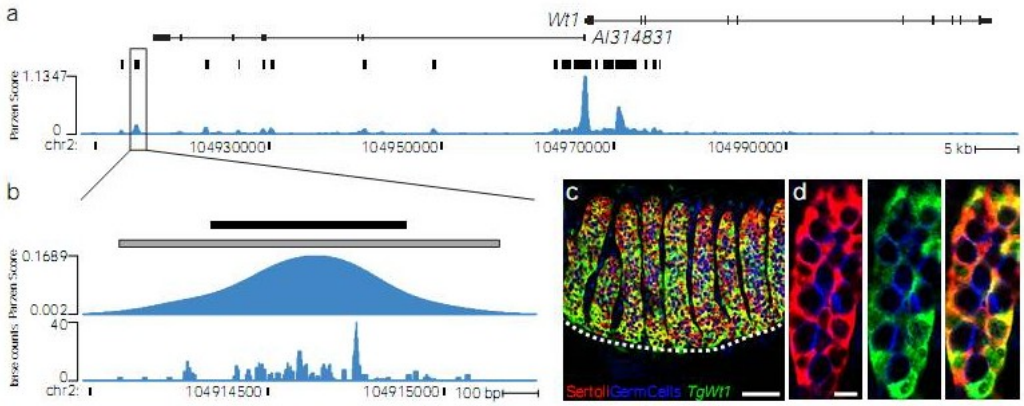


Figure 35: Identification of an enhancer of *Wt1*

(**A**) Several DHSs are observed around the *Wt1* locus. The black bars indicate the boundaries of the DHS as called by f-seq. A box shows the unique DHS picked for testing. (**B**) A close-up of the DHS picked for testing in the transient transgenic embryo. (**C**) Confocal image showing an E13.5 testis for a transgenic embryo showing expression driven by the enhancer construct. (**D**) shows the zoomed in image of a specific cord showing good overlap between Sertoli cells and reporter expression.

6.2.3 Identifying regulators from DNase-seq in E13.5 XY supporting cells

We were encouraged by our E15.5 data and went on to assay the chromatin landscape in E13.5 XY supporting cells. Our first step was to use our previous approach to identifying cell-type specific regulators of gene expression (Natarajan et al. 2012). We repeated our analysis of identifying cell-type specific regulators and assigning DHSs to them. We then scanned these regions for motifs and used a L1-sparse logistic regression classifier to classify male-enriched from female-enriched supporting cell genes. While the median AuROC on the human cells lines was ~0.7, our performance on the gonadal data was ~0.6. While this performance is clearly better than what is achieved by random (0.5), this was not a high enough classification accuracy to confidently identify regulators of gene expression in XY supporting cells.

Therefore, we took a different approach. We identified cell-type specific DHSs as we did for the E15.5 data and then set a threshold size of at least 200bp for a region to be called a putative enhancer. We repeated this analysis for the other cell types and identified putative enhancers in these cell types too. We then scanned these regions for all known motifs and used the L1-sparse logistic regression classifier. Upon doing this analysis, we found a significantly better performance in our classification with an AuROC of ~0.88. However, despite this increased performance, as before, due to high numbers of TFs belonging to same family being represented in the data, we were unable

to identify the specific TFs that were important in regulating gene expression. We then reduced the set of TFs to those that are either known to be important in sex determination and those that were identified in our fine time course expression profiling study (Munger et al. 2013). This analysis resulted in an AuROC of 0.82. Specifically, we identified that some of these TFs are likely important in sex determination and are bound in Sertoli cell specific open regions (Figure 36). Of these, *Etv4/5* is particularly interesting as it is known to be downstream of FGF signaling (Zhang et al. 2009).

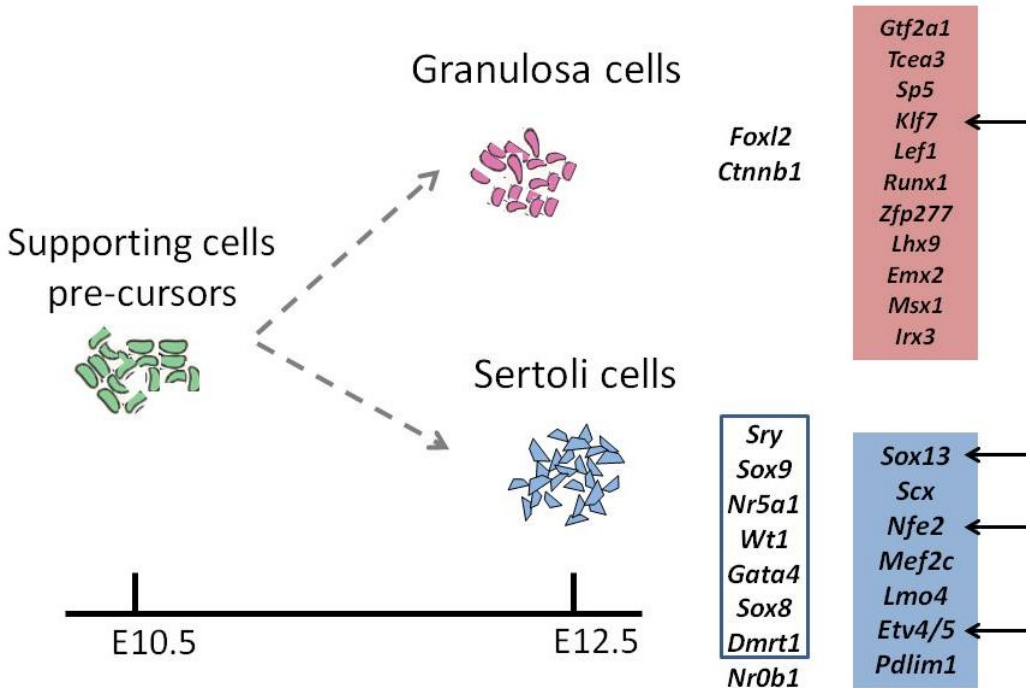


Figure 36: Factors identified by motif analysis of Sertoli specific DHS

The TFs important in sex determination from previous literature are shown in the left column of TFs. The TFs identified in our motif analysis data are outlined in the box. The TFs discovered from our study (Munger et al. 2013) are shown in the colored boxes. TFs identified in our motif analysis our indicated by an arrow.

6.2.4 ChIP-seq for H3K27ac in E13.5 XY supporting cells

To identify the active enhancers in E13.5 XY supporting cells, we attempted to conduct ChIP-seq for H3K4me1 and H3K27ac. Unfortunately, our H3K4me1 data showed high levels of background and low levels of correlation between replicates. However, our K27ac data showed high signal compared to background and a correlation of 0.89 between replicate tracks. Further, we found that >22,000 peaks were found in each replicate, with ~16,000 peaks common between replicates. Several of these peaks are around genes important in Sertoli cell differentiation. For example, at the *Gata4* locus, we saw several peak marks overlap Sertoli cell-specific DHS (Figure 37A).

We then conducted a preliminary characterization of the K27ac peaks by comparing it to the DHS peaks from multiple cell types. Crucially, we observed that the K27ac peaks are enriched specifically in DHS from Sertoli cells (Figure 37B). Further, the reciprocal analysis of analyzing the DHS from different cell types showed that while the cell-type specific DHS show enrichment for Sertoli cells, we did not see this enrichment for ubiquitous DHS peaks (Figure 37C).

We are currently conducting further analysis with the K27ac peaks to identify motifs that differentiate active enhancers from cell-type specific DHSs.

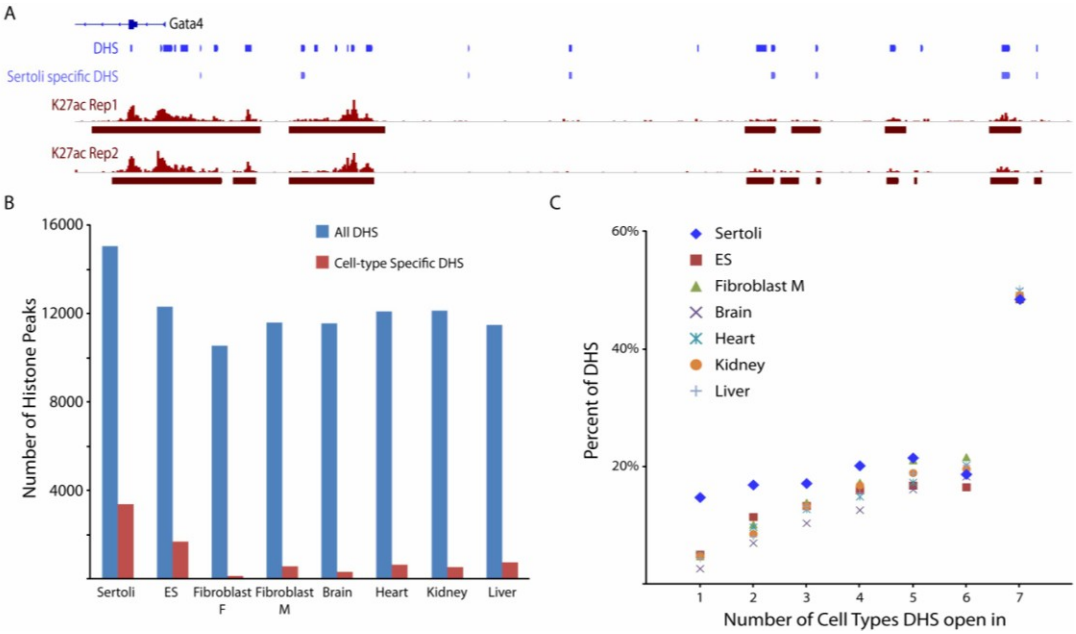


Figure 37: H3K27ac identifies active enhancers in E13.5 XY supporting cells

(**A**) Genomic locus (chr14:63,852,125-63,977,501) showing the region around *Gata4*. Several Sertoli specific peaks are observed around the gene. Further, several of these peaks overlap K27ac peaks in both replicates. (**B**) Overlap of histone peaks with DHS data from different cell lines. Sertoli-cell specific DHS have an enrichment of K27ac peak overlaps. (**C**) Based on overlap with DHS in other cell types, each DHS is classified as open in a specific number of cell types. Compared to other cell types, Sertoli-cell specific DHSs show higher overlap with the histone peak marks. However, the ubiquitous DHSs show equal overlap across all cell types.

6.3 Discussion

The work described here presents an initial attempt at characterizing the chromatin landscape in E15.5 and E13.5 XY supporting cells. While most experiments that profile genome-wide enhancers require > 10M cells, this was accomplished here with ~5M cells. This represents a significant advance and makes this approach useful in other developmental systems. Further, our initial analysis showed that by several metrics such as genomic location and size of DHS, our DNase-seq data show similar characteristics to the data from human cell lines and other cell types very well. This indicates that our attempts were indeed successful in profiling the chromatin landscape in this cell type. We also identified that the DHSs unique to Sertoli cells were enriched near Sertoli cell specific expressed genes. We used this data to identify an enhancer upstream of *Wt1* driving reporter expression. In addition, we conducted motif analysis on the Sertoli cell specific DHSs and found novel TFs that are likely involved in the differentiation process. Finally, we collected ChIP-seq data for H3K27ac and found that these peaks are enriched for overlap with Sertoli cell specific DHS.

It must be noted that our work here is still in progress and is a preliminary attempt at delving into the question of chromatin landscape in this cell type. We are currently working on conducting a more in-depth motif analysis and incorporating our ChIP-seq data set for H3K27ac into our analysis.

7. Conclusions

It is somewhat surprising and fortuitous that the different projects that I have participated in have come together in a cohesive framework.

In the first project described here, we worked with DNase-seq and expression data collected by the Crawford lab from several human cell lines (Natarajan et al. 2012). With an eye towards data that would eventually be collected from gonadal cells, we set ourselves the task of trying to predict cell-type specific expression in this wide panel of human cell lines. Crucially, this would allow us to also identify novel regulators in these cell types. We first found that the chromatin architecture at the TSS of these genes did not change significantly between cell types and if they did, it was at a specific subset of genes showing cell-type specific up-regulation. We then looked at whether sequence features in regions of open chromatin were more informative of the genes that were regulated in specific cell types than the information in the proximal promoter. Importantly, across every classification task we set, we observed far better performance when we used information from the putative distal regulatory elements identified by DNase-seq. The regression coefficients in our logistic regression model also suggested the activating or repressive roles of the TFs that were identified as informative. Our models identified known and novel TFs that likely play important regulatory roles in the

cell types we studied. Finally, at least some of these TFs showed cell-type specific footprints in the cell types where they were predicted to be functional.

In attempting a similar approach for the gonad, we first needed to collect accurate and detailed transcriptional information during the differentiation of the gonad. We conducted two experiments to perform this characterization of the differentiation process. In a work not discussed in this thesis, we analyzed expression data from different cell types in the gonad during the window of sex determination (Jameson et al. 2012b). The second project, described in this thesis, dealt with a fine time-course dataset collected from XX and XY gonads at multiple time points in a one-day interval from E11.0-E12.0 from two strains of mice (Munger et al. 2013). We designed an HMM that was specifically suited to the task of identifying cascades of transcriptional regulation. We discovered 5 cascades of regulation that generated hypotheses regarding the upstream regulators of sex determination. Interestingly, while the male-enriched genes show a strong pattern of up-regulation in XY gonads, female-enriched genes are down-regulated in the XY gonad. This indicates that in addition to an activating program, the testis has a repressive program that results in down-regulation. Further, since the ovary is, for the most part, maintaining the existing transcriptional state, it has a transcriptome that is highly similar to the bipotential gonad. This is particularly interesting as it poses questions regarding how minor transcriptional changes can

govern the commitment potential of the organ. Finally, this might also be a means for the gonad to ensure that it picks one of two different fates.

Additionally, we looked at this differentiation process in two strains of mouse – the B6 strain which is susceptible to sex reversal and the 129S1 strain which is more robust. The striking difference we observed between the transcriptomes of these two strains of mouse was the delay in the initiation of the male pathway in the B6 strain of mouse. Given the importance of timing in sex determination (Hiramatsu et al. 2009), this difference is likely to be a cause of the increased susceptibility to sex reversal in B6 mice. In addition, we integrated this data with eQTL data collected in a previous study (Munger et al. 2009) to pick candidate regulators of sex determination. We identified *Lmo4* as a good candidate and proceeded to knock it down in an *in vitro* RNAi assay with primary gonadal cells. This resulted in the down-regulation of several male-enriched genes, indicating we have discovered a novel regulator of sex determination.

Once we characterized the transcriptional differences, we collected DNase-seq data from E13.5 XY supporting cells. In addition, we developed techniques to perform ChIP-seq from low numbers of cells. We collected ChIP-seq data for H3K27ac, the histone mark correlated with active enhancer status, from XY supporting cells. Combining these two types of data we identified putative active enhancers in the XY

supporting cells. We also conducted motif analysis to identify the sequence features that are enriched in Sertoli cell specific DHSs.

8. Future Directions

It is, to my mind, very exciting that there are several possible future directions that one could take in moving forward with the work described in this thesis. While we have made significant progress in using genome-wide enhancer identification to identify important regulators in different cell types, applying these methods to uncover the TF network in mammalian sex determination, and developing techniques such as ChIP-seq with limited sample quantity, in every one of these endeavors one can see crucial improvements and future studies that need to be carried out.

8.1 Improving predictive models of gene expression

Clearly, genome-wide enhancer identification approaches such as DNase-seq, ChIP-seq and FAIRE-seq have moved us forward by leaps and bounds to understand the global chromatin configuration and they give us an extremely informative view of where TFs are bound. However, there are two additional improvements that one can foresee that will improve predictive models of gene expression.

The first of these, DNase footprinting, has been moving rapidly towards better predictions of TF binding from DNase-seq data. DNase footprinting has been used for more than 3 decades to identify sequences that are bound by TFs, but recent work has attempted to computationally glean these footprints on a genome-wide scale (Galas and Schmitz 1978; Boyle et al. 2011; Pique-Regi et al. 2011; Neph et al. 2012). Specifically, while some factors such as NRSF and CTCF form well defined footprints, other factors are not so pronounced in their footprints. Furthermore, DNaseI seems to have a limited, but not insignificant, sequence bias that has to be accounted for when one tries to identify binding sites of TFs (Koohy et al. 2013; He et al. 2014). This is particularly problematic when the TF in question has a motif that is affected by the DNaseI sequence bias. Nonetheless, recent techniques, including those being developed in the Ohler lab, are beginning to account for this and improve the accuracy with which one can identify

occupied TFBSs in DNase-seq peaks (He et al. 2014; Sherwood et al. 2014). Progress in this front will improve predictive models of gene expression.

Another key improvement is likely to be in identifying the interactions between specific enhancers and genes. Given that enhancers do not interact with genes in a manner determined by linear distance, our use of proximity as a proxy for interaction is an over-simplification (Sanyal et al. 2012; Jin et al. 2013). The experimental techniques most pertinent to addressing this issue are the suite of chromosome conformation capture techniques such as 3C, 4C, 5C and Hi-C (de Wit and de Laat 2012). These techniques are typically limited by their lack of resolution, although one recent study reported a resolution of approximately 5-10kb (Jin et al. 2013). This level of resolution will be particularly useful in understanding which enhancers interact with which promoters. Another experimental approach that is beneficial to identify enhancer promoter interactions is ChIA-PET or Chromatin Interaction Analysis by Paired-End Tag sequencing, which identifies proteins that are concurrently proximal to multiple regions of the genome (Fullwood et al. 2009). This technique was recently used with a PolII IP, resulting in a high resolution interactome map of the transcribed regions of the genome (Kieffer-Kwon et al. 2013). Similar assays can be carried out under conditions of environmental stimuli, for example, to understand the dynamics of chromosome conformation during gene expression changes.

Alternatively, computational and data analysis techniques can be used to identify putative interactions. Recent work has used DNase-seq data across several cell lines, to identify cell-type specific DHSs with expression changes of genes in *cis* (Sheffield et al. 2013). Similar approaches have been used with data from chromatin marks and PolII ChIP (Akhtar-Zaidi et al. 2012; Shen et al. 2012). While one can identify several enhancer-promoter interactions in this way, the need for data from several cell lines makes this a less than ideal approach. Recent work in the Ohler lab, attempts to circumvent this by jointly learning both the factors that regulate gene expression and the regions of the genome where these TFs bind near the genes that show cell-type specific expression patterns. This is a promising approach that might reveal the principles of enhancer-promoter interactions.

Taken from a longer perspective, the problem of how enhancers communicate with promoters is one of central importance in mammalian gene regulation (Krivega and Dean 2012). Therefore, one might want to step back and instead of trying to predict gene expression, attempt to predict the chromosome conformation gleaned from experimental assays. Recent data indicates that topological domains are organized by CTCF, cohesin and lamins (Lund et al. 2013; Baranello et al. 2014; Collas et al. 2014). Using this data, and associated chromatin marks (H3K9me3, in particular) to model

chromosome conformation would be an extremely useful advance to understanding gene regulation.

8.2 Transcriptional changes and their causes in the supporting cells

Our work here has helped us move significantly closer to understanding the transcriptional behavior and the factors at play in supporting cells, particularly XY supporting cells. However, it has also generated several immediate questions.

First among these is which factors mediate the down-regulation of the female-enriched genes in the male gonad. While we knew that a few genes were down-regulated, we did not envision that hundreds of genes would show this trend. It remains to be seen whether all these genes are each independently repressed by a large network, or if only an activator of female-enriched genes is actively repressed, resulting in the consequent down-regulation of several of these genes. One hypothesis is that this repression is mediated by the polycomb complex. This is particularly interesting because the polycomb component, *Cbx2*, is known to be important for male development (Katoh-Fukui et al. 1998; Katoh-Fukui et al. 2012). Current experiments in the Capel lab are focused on identifying the binding locations of the polycomb factor, *Ring1b*, in XY supporting cells. These experiments could be highly informative about the mechanism of the repression. However, it must be kept in mind that while polycomb is in general an attractive candidate for repression, it must be recruited to a specific locus (Schuettengruber and Cavalli 2009). Therefore, even if polycomb is involved in the

repression of female-enriched genes, key questions will still remain about the factor that provides the specificity for these genes to be repressed.

Second, while we successfully identified the onset of the male program, our time-course of expression did not identify the onset of the genes in the bipotential gonad. This is particularly important given that the early differentiated ovary (E12.0) is very similar to the bipotential gonad. Therefore, we are not only missing out on the transcriptional cascades that give rise to the bipotential gonad, but also the genes that are eventually female-enriched. One way to approach this problem is to profile the expression at a fine time granularity at earlier stages (~E10.5). This can be done in a sorted cell population from cells that can be obtained from the *Sf1-EGFP* mouse (Nef et al. 2005). These cells are expected to be, for the most part, bipotential supporting cells that will eventually express SRY. This approach could also be informative regarding the transcriptional dynamics in the gonad prior to the expression of *Sry*. It is important to note that, apart from genes on the Y-chromosome, and some genes on the X-chromosome, there will be no expression differences between XX and XY gonads. Therefore, to identify cascades, the HMM would have to be tweaked to monitor temporal changes between adjacent time points.

Third, a major question opened up by previous studies is how the critical window of sex determination is kept open for the initiation of the male pathway

(Hiramatsu et al. 2009). Specifically, it has been noticed that the timing of *Sry* activation is tightly regulated to a 6 hour time period, which if missed, will result in ovarian development. Our time course transcriptome, while suggesting possible genes involved in implementing this behavior, also adds a level of intrigue to it. Namely, given that the transcriptome of the bipotential gonad and the ovary are highly similar, how do the minor changes observed allow the bipotential gonad to activate the male program while impairing this ability in the early differentiated ovary. Of the few genes that show dynamic changes in expression, a TF that is strongly up-regulated in XX gonads at E11.6 is *Irx3* (Munger et al. 2013). While the mutant for this gene has not shown a striking embryonic phenotype (Jorgensen and Gao 2005; Kim et al. 2011), I would argue that the experiments conducted do not test the specific hypothesis that *Irx3* is the gene that regulates the window of male pathway activation by *Sry*. Two experiments can be conducted to test this hypothesis. One experiment would entail using an XX *Irx3*$^{-/-}$ embryo that also has an *Hsp-Sry* transgene (Hiramatsu et al. 2009). If the hypothesis that *Irx3* regulates the window of sex determination is true, it follows that removing *Irx3* should extend the window of potential male pathway activation. Therefore, delayed activation of *Sry* by heat shock should still result in testis development. Alternatively, one could attempt to ectopically express *Irx3* in XY gonads and observe if the male pathway can be initiated.

8.3 The TF network in multiple cell types in the gonad

The majority of our success has been in understanding the transcriptional dynamics and their causes in the XY supporting cells of the gonad. Due to technical difficulties, including obtaining sufficient biological material, we were unable to glean similar information from XX supporting cells. For similar reasons, we did not attempt to collect data from the bipotential supporting cell precursors. An understanding of these two cell types will be crucial to improving our knowledge of how chromatin can allow for potential in precursor cells and then proceed to shut down this potential as differentiation progresses.

There are good reasons to hope that the technical difficulties will not be insurmountable in the near future. Chief among these reasons, is the development of a new protocol named ATAC-seq, or Assay for Transposase-Accessible Chromatin using sequencing (Buenrostro et al. 2013). In this technique, fresh cells are collected, nuclei are isolated, following which a transposase is introduced. This transposase, which can only access regions of open chromatin, inserts an adapter at these sites which are then used to build a library. The protocol has been optimized to work at ~50,000 cells, two orders of magnitude less than the 5 million cells needed for DNase-seq. This technological advance opens the door to understanding the chromatin landscape of the XX differentiated cells and the early bipotential cells. As for improvements in ChIP-seq, we

tried to perform ChIP-seq with ~40,000 cells/IP with XX supporting cells, but were not successful. Given this was a first attempt, if resources are invested to optimize the small cell number assay, we will likely be able to obtain useful and illuminating data in these cell types.

Another level of detail that we have ignored in our study is that supporting cells are derived from cells in the coelomic epithelium (Karl and Capel 1998). Interestingly, these cells seem to be the precursor cells for at least one other cell type, called the interstitial cells. Therefore, in addition to the sexual fate decision that we have been concerned with, the supporting cells have had to already differentiate themselves from the interstitial cells. It has been hard to penetrate how they make this decision, due to the difficulties in sorting the cells of the coelomic epithelium. Nonetheless, we can sort supporting cells and interstitial cells in the E13.5 gonad and look for differences in the chromatin between these two cell types. This too would be an effective and potentially less challenging means of understanding how cell fate decisions are made.

The supporting cells serve as a fantastic model for bipotential cells which arise in a variety of developmental systems. However, the unique property of the gonad is that the whole organ is bipotential, with multiple cell types concurrently committing to one of two different fates. This situation is unique in organogenesis and the bipotentiality of the early gonad primordium is conserved across vertebrates (Barske and Capel 2008).

How such a decision is implemented in the chromatin of multiple cells is a fascinating question that can only be approached by research in the gonad. The mouse gonad is ideally suited to this problem, due to the development of multiple transgenic lines of mice that enable the sorting of distinct cell populations at various stages of their development. One can study how the existing chromatin environment allows the germ cells and interstitial cells to respond to the signaling from the supporting cells. The motifs in the cell-type specifically enriched enhancers will likely indicate which extracellular signaling pathways are operative in each of these cell types. Comparisons between XX and XY cells will reveal the sex-specific signals responsible for the sexual dimorphism of these cell types. If I had a few more years in graduate school, these fascinating questions would fully occupy my time.

References

Adli M, Bernstein BE. 2011. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nat Protoc* **6**(10): 1656-1668.

Adli M, Zhu J, Bernstein BE. 2010. Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. *Nat Methods* **7**(8): 615-618.

Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B. 2003. Computational detection of cis -regulatory modules. *Bioinformatics* **19 Suppl 2**: ii5-14.

Akhtar-Zaidi B, Cowper-Sal-lari R, Corradin O, Saiakhova A, Bartels CF, Balasubramanian D, Myeroff L, Lutterbaugh J, Jarrar A, Kalady MF et al. 2012. Epigenomic enhancer profiling defines a signature of colon cancer. *Science* **336**(6082): 736-739.

Andersson R Gebhard C Miguel-Escalada I Hoof I Bornholdt J Boyd M Chen Y Zhao X Schmidl C Suzuki T et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**(7493): 455-461.

Asp P, Blum R, Vethantham V, Parisi F, Micsinai M, Cheng J, Bowman C, Kluger Y, Dynlacht BD. 2011. Genome-wide remodeling of the epigenetic landscape during myogenic differentiation. *Proc Natl Acad Sci U S A* **108**(22): E149-158.

Asprer JS, Lee B, Wu CS, Vadakkan T, Dickinson ME, Lu HC, Lee SK. 2011. LMO4 functions as a co-activator of neurogenin 2 in the developing cortex. *Development* **138**(13): 2823-2832.

Bailey TL, Boden M, Whitington T, Machanick P. 2010. The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics* **11**.

Banerji J, Rusconi S, Schaffner W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**(2 Pt 1): 299-308.

Baranello L, Kouzine F, Levens D. 2014. CTCF and cohesin cooperate to organize the 3D structure of the mammalian genome. *Proc Natl Acad Sci U S A* **111**(3): 889-890.

Barske LA, Capel B. 2008. Blurring the edges in vertebrate sex determination. *Curr Opin Genet Dev* **18**(6): 499-505.

Beer MA, Tavazoie S. 2004. Predicting gene expression from sequence. *Cell* **117**(2): 185-198.

Bemmo A, Benovoy D, Kwan T, Gaffney DJ, Jensen RV, Majewski J. 2008. Gene expression and isoform variation analysis using Affymetrix Exon Arrays. *BMC Genomics* **9**: 529.

Beverdam A, Koopman P. 2006. Expression profiling of purified mouse gonadal somatic cells during the critical time window of sex determination reveals novel candidate genes for human sexual dysgenesis syndromes. *Hum Mol Genet* **15**(3): 417-431.

Bishop CE, Whitworth DJ, Qin Y, Agoulnik AI, Agoulnik IU, Harrison WR, Behringer RR, Overbeek PA. 2000. A transgenic insertion upstream of sox9 is associated with dominant XX sex reversal in the mouse. *Nat Genet* **26**(4): 490-494.

Bitgood MJ, Shen L, McMahon AP. 1996. Sertoli cell signaling by Desert hedgehog regulates the male germline. *Curr Biol* **6**(3): 298-304.

Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F et al. 2010. ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* **42**(9): 806-810.

Blum R, Vethantham V, Bowman C, Rudnicki M, Dynlacht BD. 2012. Genome-wide identification of enhancers in skeletal muscle: the role of MyoD1. *Genes Dev* **26**(24): 2763-2779.

Boehm AK, Saunders A, Werner J, Lis JT. 2003. Transcription factor and polymerase recruitment, modification, and movement on dhsp70 in vivo in the minutes following heat shock. *Mol Cell Biol* **23**(21): 7628-7637.

Bouma GJ, Affourtit JP, Bult CJ, Eicher EM. 2007. Transcriptional profile of mouse pre-granulosa and Sertoli cells isolated from early-differentiated fetal gonads. *Gene Expr Patterns* **7**(1-2): 113-123.

Bouma GJ, Albrecht KH, Washburn LL, Recknagel AK, Churchill GA, Eicher EM. 2005. Gonadal sex reversal in mutant Dax1 XY mice: a failure to upregulate Sox9 in pre-Sertoli cells. *Development* **132**(13): 3045-3054.

Bouma GJ, Hudson QJ, Washburn LL, Eicher EM. 2010. New candidate genes identified for controlling mouse gonadal sex determination and the early stages of granulosa and Sertoli cell differentiation. *Biol Reprod* **82**(2): 380-389.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008a. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**(2): 311-322.

Boyle AP, Guinney J, Crawford GE, Furey TS. 2008b. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**(21): 2537-2538.

Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS. 2011. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res* **21**(3): 456-464.

Bradford ST, Hiramatsu R, Maddugoda MP, Bernard P, Chaboissier MC, Sinclair A, Schedl A, Harley V, Kanai Y, Koopman P et al. 2009. The cerebellin 4 precursor gene is a direct target of SRY and SOX9 in mice. *Biol Reprod* **80**(6): 1178-1188.

Brennan J, Capel B. 2004. One tissue, two fates: molecular genetic events that underlie testis versus ovary development. *Nat Rev Genet* **5**(7): 509-521.

Brush SG. 1978. Nettie M. Stevens and the discovery of sex determination by chromosomes. *Isis* **69**(247): 163-172.

Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A. 2008. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res* **36**(Database issue): D102-106.

Buecker C, Wysocka J. 2012. Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet* **28**(6): 276-284.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**(12): 1213-1218.

Bulger M, Groudine M. 2011. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**(3): 327-339.

Bullejos M, Koopman P. 2005. Delayed Sry and Sox9 expression in developing mouse gonads underlies B6-Y(DOM) sex reversal. *Dev Biol* **278**(2): 473-481.

Bunt J, Hasselt NE, Zwijnenburg DA, Hamdi M, Koster J, Versteeg R, Kool M. 2011. OTX2 directly activates cell cycle genes and inhibits differentiation in medulloblastoma cells. *Int J Cancer*.

Carlson EA. 2013. *The 7 Sexes: Biology of Sex Determination*. Indiana University Press.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**(6): 626-635.

Cereghini S. 1996. Liver-enriched transcription factors and hepatocyte differentiation. *FASEB J* **10**(2): 267-282.

Chen T, Dent SY. 2014. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat Rev Genet* **15**(2): 93-106.

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J et al. 2008. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**(6): 1106-1117.

Chuva de Sousa Lopes SM, Hayashi K, Shovlin TC, Mifsud W, Surani MA, McLaren A. 2008. X chromosome activity in mouse XX primordial germ cells. *PLoS Genet* **4**(2): e30.

Collas P, Lund EG, Oldenburg AR. 2014. Closing the (nuclear) envelope on the genome: how nuclear lamins interact with promoters and modulate gene expression. *Bioessays* **36**(1): 75-83.

Colvin JS, Green RP, Schmahl J, Capel B, Ornitz DM. 2001. Male-to-female sex reversal in mice lacking fibroblast growth factor 9. *Cell* **104**(6): 875-889.

Correa SM, Washburn LL, Kahlon RS, Musson MC, Bouma GJ, Eicher EM, Albrecht KH. 2012. Sex reversal in C57BL/6J XY mice caused by increased expression of ovarian genes and insufficient activation of the testis determining pathway. *PLoS Genet* **8**(4): e1002569.

Cowper-Sal lari R, Zhang X, Wright JB, Bailey SD, Cole MD, Eeckhoute J, Moore JH, Lupien M. 2012. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat Genet* **44**(11): 1191-1198.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**(50): 21931-21936.

Cuddapah S, Barski A, Cui K, Schones DE, Wang Z, Wei G, Zhao K. 2009. Native chromatin preparation and Illumina/Solexa library construction. *Cold Spring Harb Protoc* **2009**(6): pdb prot5237.

Dahl JA, Collas P. 2009. MicroChIP: chromatin immunoprecipitation for small cell numbers. *Methods Mol Biol* **567**: 59-74.

Darnell J. 2011. *RNA: Life's Indispensable Molecule*. Cold Spring Harbor Laboratory Press, New York.

Das D, Nahle Z, Zhang MQ. 2006. Adaptively inferring human transcriptional subnetworks. *Mol Syst Biol* **2**: 2006 0029.

De Santa Barbara P, Bonneaud N, Boizet B, Desclozeaux M, Moniot B, Sudbeck P, Scherer G, Poulat F, Berta P. 1998. Direct interaction of SRY-related protein SOX9 and steroidogenic factor 1 regulates transcription of the human anti-Mullerian hormone gene. *Mol Cell Biol* **18**(11): 6653-6665.

de Wit E, de Laat W. 2012. A decade of 3C technologies: insights into nuclear organization. *Genes Dev* **26**(1): 11-24.

DeAngelis MM, Wang DG, Hawkins TL. 1995. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res* **23**(22): 4742-4743.

DeJong J, Bernstein R, Roeder RG. 1995. Human general transcription factor TFIIA: characterization of a cDNA encoding the small subunit and requirement for basal and activated transcription. *Proc Natl Acad Sci U S A* **92**(8): 3313-3317.

Deng W, Lee J, Wang H, Miller J, Reik A, Gregory PD, Dean A, Blobel GA. 2012. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**(6): 1233-1244.

Dy P, Wang W, Bhattaram P, Wang Q, Wang L, Ballock RT, Lefebvre V. 2012. Sox9 directs hypertrophic maturation and blocks osteoblast differentiation of growth plate chondrocytes. *Dev Cell* **22**(3): 597-609.

Eggers S, Sinclair A. 2012. Mammalian sex determination-insights from humans and mice. *Chromosome Res*.

Eicher EM, Washburn LL. 1983. Inherited sex reversal in mice: identification of a new primary sex-determining gene. *J Exp Zool* **228**(2): 297-304.

Eicher EM, Washburn LL, Schork NJ, Lee BK, Shown EP, Xu X, Dredge RD, Pringle MJ, Page DC. 1996. Sex-determining genes on mouse autosomes identified by linkage analysis of C57BL/6J-YPOS sex reversal. *Nat Genet* **14**(2): 206-209.

Eicher EM, Washburn LL, Whitney JB, 3rd, Morrow KE. 1982. Mus poschiavinus Y chromosome in the C57BL/6J murine genome causes sex reversal. *Science* **217**(4559): 535-537.

Elnitski L, Jin VX, Farnham PJ, Jones SJ. 2006. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res* **16**(12): 1455-1464.

Fan H, Cinar MU, Phatsara C, Tesfaye D, Tholen E, Looft C, Schellander K. 2011. Molecular mechanism underlying the differential MYF6 expression in postnatal skeletal muscle of Duroc and Pietrain breeds. *Gene* **486**(1-2): 8-14.

Ford CE, Jones KW, Polani PE, De Almeida JC, Briggs JH. 1959. A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). *Lancet* **1**(7075): 711-713.

Fu Y, Frith MC, Haverty PM, Weng Z. 2004. MotifViz: an analysis and visualization tool for motif discovery. *Nucleic Acids Res* **32**(Web Server issue): W420-423.

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH et al. 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**(7269): 58-64.

Furey TS. 2012. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet* **13**(12): 840-852.

Galas DJ, Schmitz A. 1978. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* **5**(9): 3157-3170.

Garcia-Ortiz JE, Pelosi E, Omari S, Nedorezov T, Piao Y, Karmazin J, Uda M, Cao A, Cole SW, Forabosco A et al. 2009. Foxl2 functions in sex determination and histogenesis throughout mouse ovary development. *BMC Dev Biol* **9**: 36.

Gilbert W, Muller-Hill B. 1966. Isolation of the lac repressor. *Proc Natl Acad Sci U S A* **56**(6): 1891-1898.

Gotea V, Ovcharenko I. 2008. DiRE: identifying distant regulatory elements of co-expressed genes. *Nucleic Acids Res* **36**(Web Server issue): W133-139.

Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res* **20**(5): 565-577.

Gubbay J, Collignon J, Koopman P, Capel B, Economou A, Munsterberg A, Vivian N, Goodfellow P, Lovell-Badge R. 1990. A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. *Nature* **346**(6281): 245-250.

Haeussler M, Joly JS. 2011. When needles look like hay: how to find tissue-specific enhancers in model organism genomes. *Dev Biol* **350**(2): 239-254.

Hahm K, Sum EY, Fujiwara Y, Lindeman GJ, Visvader JE, Orkin SH. 2004. Defective neural tube closure and anteroposterior patterning in mice lacking the LIM protein LMO4 or its interacting partner Deaf-1. *Mol Cell Biol* **24**(5): 2074-2082.

Hampsey M. 1998. Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol Mol Biol Rev* **62**(2): 465-503.

Haring M, Offermann S, Danker T, Horst I, Peterhansel C, Stam M. 2007. Chromatin immunoprecipitation: optimization, quantitative analysis and data normalization. *Plant methods* **3**: 11.

He HH, Meyer CA, Hu SS, Chen MW, Zang C, Liu Y, Rao PK, Fei T, Xu H, Long H et al. 2014. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat Methods* **11**(1): 73-78.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**(7243): 108-112.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**(3): 311-318.

Henikoff S, Shilatifard A. 2011. Histone modification: cause or cog? *Trends Genet* **27**(10): 389-396.

Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS et al. 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* **6**(4): 283-289.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**(23): 9362-9367.

Hiramatsu R, Matoba S, Kanai-Azuma M, Tsunekawa N, Katoh-Fukui Y, Kurohmaru M, Morohashi K, Wilhelm D, Koopman P, Kanai Y. 2009. A critical time window of Sry action in gonadal sex determination in mice. *Development* **136**(1): 129-138.

Huang DW, Sherman BT, Lempicki RA. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37**(1): 1-13.

-. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**(1): 44-57.

Huang DY, Kuo YY, Chang ZF. 2005. GATA-1 mediates auto-regulation of Gfi-1B transcription in K562 cells. *Nucleic Acids Res* **33**(16): 5331-5342.

Jacobs PA, Strong JA. 1959. A case of human intersexuality having a possible XXY sex-determining mechanism. *Nature* **183**(4657): 302-303.

Jameson SA, Lin YT, Capel B. 2012a. Testis development requires the repression of Wnt4 by Fgf signaling. *Dev Biol* **370**(1): 24-32.

Jameson SA, Natarajan A, Cool J, DeFalco T, Maatouk DM, Mork L, Munger SC, Capel B. 2012b. Temporal transcriptional profiling of somatic and germ cells reveals biased lineage priming of sexual fate in the fetal mouse gonad. *PLoS Genet* **8**(3): e1002575.

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, Schmitt AD, Espinoza CA, Ren B. 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**(7475): 290-294.

John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA. 2011. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* **43**(3): 264-268.

Johnson WE, Li C, Rabinovic A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**(1): 118-127.

Jorgensen JS, Gao L. 2005. Irx3 is differentially up-regulated in female gonads during sex determination. *Gene Expr Patterns* **5**(6): 756-762.

Josso N. 2008. Professor Alfred Jost: the builder of modern sex differentiation. *Sex Dev* **2**(2): 55-63.

Kanazawa S, Soucek L, Evan G, Okamoto T, Peterlin BM. 2003. c-Myc recruits P-TEFb for transcription, cellular proliferation and apoptosis. *Oncogene* **22**(36): 5707-5711.

Kantorovitz MR, Kazemian M, Kinston S, Miranda-Saavedra D, Zhu Q, Robinson GE, Gottgens B, Halfon MS, Sinha S. 2009. Motif-blind, genome-wide discovery of cis-regulatory modules in Drosophila and mouse. *Dev Cell* **17**(4): 568-579.

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**(7236): 362-366.

Karl J, Capel B. 1998. Sertoli cells of the mouse testis originate from the coelomic epithelium. *Dev Biol* **203**(2): 323-333.

Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. 2010. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A* **107**(7): 2926-2931.

Katoh-Fukui Y, Miyabayashi K, Komatsu T, Owaki A, Baba T, Shima Y, Kidokoro T, Kanai Y, Schedl A, Wilhelm D et al. 2012. Cbx2, a polycomb group gene, is required for Sry gene expression in mice. *Endocrinology* **153**(2): 913-924.

Katoh-Fukui Y, Tsuchiya R, Shiroishi T, Nakahara Y, Hashimoto N, Noguchi K, Higashinakagawa T. 1998. Male-to-female sex reversal in M33 mutant mice. *Nature* **393**(6686): 688-692.

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**(7364): 289-294.

Kieffer-Kwon KR, Tang Z, Mathe E, Qian J, Sung MH, Li G, Resch W, Baek S, Pruett N, Grontved L et al. 2013. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell* **155**(7): 1507-1520.

Kim B, Kim Y, Cooke PS, Ruther U, Jorgensen JS. 2011. The fused toes locus is essential for somatic-germ cell interactions that foster germ cell maturation in developing gonads in mice. *Biol Reprod* **84**(5): 1024-1032.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**(7295): 182-187.

Kim Y, Bingham N, Sekido R, Parker KL, Lovell-Badge R, Capel B. 2007. Fibroblast growth factor receptor 2 regulates proliferation and Sertoli differentiation during male sex determination. *Proc Natl Acad Sci U S A* **104**(42): 16558-16563.

Kim Y, Kobayashi A, Sekido R, DiNapoli L, Brennan J, Chaboissier MC, Poulat F, Behringer RR, Lovell-Badge R, Capel B. 2006. Fgf9 and Wnt4 act as antagonistic signals to regulate mammalian sex determination. *PLoS Biol* **4**(6): e187.

Koh K, Kim S-J, Boyd S. 2007. An interior-point method for largescale l1-regularized logistic regression. *Journal of Machine Learning Research* **8**: 1519-1555.

Kohwi M, Lupton JR, Lai SL, Miller MR, Doe CQ. 2013. Developmentally regulated subnuclear genome reorganization restricts neural progenitor competence in Drosophila. *Cell* **152**(1-2): 97-108.

Koohy H, Down TA, Hubbard TJ. 2013. Chromatin accessibility data sets show bias due to sequence specificity of the DNase I enzyme. *PLoS One* **8**(7): e69853.

Kool M, Koster J, Bunt J, Hasselt NE, Lakeman A, van Sluis P, Troost D, Meeteren NS, Caron HN, Cloos J et al. 2008. Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. *PLoS One* **3**(8): e3088.

Koopman P, Munsterberg A, Capel B, Vivian N, Lovell-Badge R. 1990. Expression of a candidate sex-determining gene during mouse testis differentiation. *Nature* **348**(6300): 450-452.

Krivega I, Dean A. 2012. Enhancer and promoter interactions-long distance calls. *Curr Opin Genet Dev* **22**(2): 79-85.

Landolin JM, Johnson DS, Trinklein ND, Aldred SF, Medina C, Shulha H, Weng Z, Myers RM. 2010. Sequence features that drive human promoter function and tissue specificity. *Genome Res* **20**(7): 890-898.

Lee SK, Jurata LW, Nowak R, Lettieri K, Kenny DA, Pfaff SL, Gill GN. 2005. The LIM domain-only protein LMO4 is required for neural tube closure. *Mol Cell Neurosci* **28**(2): 205-214.

Lercher MJ, Urrutia AO, Pavlicek A, Hurst LD. 2003. A unification of mosaic structures in the human genome. *Hum Mol Genet* **12**(19): 2411-2415.

Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**(14): 1725-1735.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Liu H, Schmidt-Supprian M, Shi Y, Hobeika E, Barteneva N, Jumaa H, Pelanda R, Reth M, Skok J, Rajewsky K. 2007. Yin Yang 1 is a critical regulator of B-cell development. *Genes Dev* **21**(10): 1179-1189.

Lodato MA, Ng CW, Wamstad JA, Cheng AW, Thai KK, Fraenkel E, Jaenisch R, Boyer LA. 2013. SOX2 co-occupies distal enhancer elements with distinct POU factors in ESCs and NPCs to specify cell state. *PLoS Genet* **9**(2): e1003288.

Lovell-Badge R, Robertson E. 1990. XY female mice resulting from a heritable mutation in the primary testis-determining gene, Tdy. *Development* **109**(3): 635-646.

Lu R, Medina KL, Lancki DW, Singh H. 2003. IRF-4,8 orchestrate the pre-B-to-B transition in lymphocyte development. *Genes Dev* **17**(14): 1703-1708.

Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M. 2005. Functional evolution of a cis-regulatory module. *PLoS Biol* **3**(4): e93.

Lund E, Oldenburg AR, Delbarre E, Freberg CT, Duband-Goulet I, Eskeland R, Buendia B, Collas P. 2013. Lamin A/C-promoter interactions specify chromatin state-dependent transcription outcomes. *Genome Res* **23**(10): 1580-1589.

Maatouk DM, DiNapoli L, Alvers A, Parker KL, Taketo MM, Capel B. 2008. Stabilization of beta-catenin in XY gonads causes male-to-female sex-reversal. *Hum Mol Genet* **17**(19): 2949-2955.

Mahony S, Benos PV. 2007. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* **35**(Web Server issue): W253-258.

Marfil V, Moya M, Pierreux CE, Castell JV, Lemaigre FP, Real FX, Bort R. 2010. Interaction between Hhex and SOX13 modulates Wnt/TCF activity. *J Biol Chem* **285**(8): 5726-5737.

Margaritis T, Oreal V, Brabers N, Maestroni L, Vitaliano-Prunier A, Benschop JJ, van Hooff S, van Leenen D, Dargemont C, Geli V et al. 2012. Two distinct repressive mechanisms for histone 3 lysine 4 methylation through promoting 3'-end antisense transcription. *PLoS Genet* **8**(9): e1002952.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K et al. 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**(Database issue): D108-110.

McKay DJ, Lieb JD. 2013. A common set of DNA regulatory elements shapes Drosophila appendages. *Dev Cell* **27**(3): 306-318.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**(5): 495-501.

Megraw M, Pereira F, Jensen ST, Ohler U, Hatzigeorgiou AG. 2009. A transcription factor affinity-based code for mammalian transcription initiation. *Genome Res* **19**(4): 644-656.

Melichar HJ, Narayan K, Der SD, Hiraoka Y, Gardiol N, Jeannet G, Held W, Chambers CA, Kang J. 2007. Regulation of gammadelta versus alphabeta T lymphocyte differentiation by the transcription factor SOX13. *Science* **315**(5809): 230-233.

Menke DB, Page DC. 2002. Sexually dimorphic gene expression in the developing mouse gonad. *Gene Expr Patterns* **2**(3-4): 359-367.

Michell AC, Braganca J, Broadbent C, Joyce B, Franklyn A, Schneider JE, Bhattacharya S, Bamforth SD. 2010. A novel role for transcription factor Lmo4 in thymus development through genetic interaction with Cited2. *Dev Dyn* **239**(7): 1988-1994.

Moffat J, Grueneberg DA, Yang X, Kim SY, Kloepfer AM, Hinkle G, Piqani B, Eisenhaure TM, Luo B, Grenier JK et al. 2006. A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* **124**(6): 1283-1298.

Morais da Silva S, Hacker A, Harley V, Goodfellow P, Swain A, Lovell-Badge R. 1996. Sox9 expression during gonadal development implies a conserved role for the gene in testis differentiation in mammals and birds. *Nat Genet* **14**(1): 62-68.

Morgan TH. 1903. Recent theories in regard to the determination of sex. *The Popular Science Monthly* **64**.

Munger SC, Aylor DL, Syed HA, Magwene PM, Threadgill DW, Capel B. 2009. Elucidation of the transcription network governing mammalian sex determination by exploiting strain-specific susceptibility to sex reversal. *Genes Dev* **23**(21): 2521-2536.

Munger SC, Natarajan A, Looger LL, Ohler U, Capel B. 2013. Fine time course expression analysis identifies cascades of activation and repression and maps a putative regulator of mammalian sex determination. *PLoS Genet* **9**(7): e1003630.

Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U. 2012. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res* **22**(9): 1711-1722.

Nef S, Schaad O, Stallings NR, Cederroth CR, Pitetti JL, Schaer G, Malki S, Dubois-Dauphin M, Boizet-Bonhoure B, Descombes P et al. 2005. Gene expression during sex determination reveals a robust female genetic program at the onset of ovarian development. *Dev Biol* **287**(2): 361-377.

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK et al. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**(7414): 83-90.

Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res* **37**(Database issue): D77-82.

Ng JH, Kumar V, Muratani M, Kraus P, Yeo JC, Yaw LP, Xue K, Lufkin T, Prabhakar S, Ng HH. 2013. In vivo epigenomic profiling of germ cells reveals germ cell molecular signatures. *Dev Cell* **24**(3): 324-333.

Nikolova G, Sinsheimer JS, Eicher EM, Vilain E. 2008. The chromosome 11 region from strain 129 provides protection from sex reversal in XYPOS mice. *Genetics* **179**(1): 419-427.

Nord AS, Blow MJ, Attanasio C, Akiyama JA, Holt A, Hosseini R, Phouanenavong S, Plajzer-Frick I, Shoukry M, Afzal V et al. 2013. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155**(7): 1521-1531.

O'Neill LP, VerMilyea MD, Turner BM. 2006. Epigenetic characterization of the early embryo with a chromatin immunoprecipitation protocol applicable to small cell populations. *Nature Genetics* **38**(7): 835-841.

Ochoa SD, Salvador S, LaBonne C. 2012. The LIM adaptor protein LMO4 is an essential regulator of neural crest development. *Dev Biol* **361**(2): 313-325.

Oda N, Abe M, Sato Y. 1999. ETS-1 converts endothelial cells to the angiogenic phenotype by inducing the expression of matrix metalloproteinases and integrin beta3. *J Cell Physiol* **178**(2): 121-132.

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39**(6): 730-732.

Ogryzko VV, Schiltz RL, Russanova V, Howard BH, Nakatani Y. 1996. The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell* **87**(5): 953-959.

Ong CT, Corces VG. 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics* **12**(4): 283-293.

Ouyang Z, Zhou Q, Wong WH. 2009. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci U S A* **106**(51): 21521-21526.

Painter TS. 1923. Studies in mammalian spermatogenesis II: The spermatogenesis of man. *Journal of Experimental Zoology* **37**: 291–336.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**(7118): 499-502.

Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. 2011. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* **21**(3): 447-455.

Pitetti JL, Calvel P, Romero Y, Conne B, Truong V, Papaioannou MD, Schaad O, Docquier M, Herrera PL, Wilhelm D et al. 2013. Insulin and IGF1 receptors are

essential for XX and XY gonadal differentiation and adrenal development in mice. *PLoS Genet* **9**(1): e1003160.

Ptashne M. 1967. ISOLATION OF THE lambda PHAGE REPRESSOR. *Proc Natl Acad Sci U S A* **57**(2): 306-313.

-. 1986. Gene regulation by proteins acting nearby and at a distance. *Nature* **322**(6081): 697-701.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6): 841-842.

Rada-Iglesias A, Bajpai R, Prescott S, Brugmann SA, Swigut T, Wysocka J. 2012. Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. *Cell Stem Cell* **11**(5): 633-648.

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**(7333): 279-283.

Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA. 2010. c-Myc regulates transcriptional pause release. *Cell* **141**(3): 432-445.

Ramsey SA, Klemm SL, Zak DE, Kennedy KA, Thorsson V, Li B, Gilchrist M, Gold ES, Johnson CD, Litvak V et al. 2008. Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics. *PLoS Comput Biol* **4**(3): e1000021.

Ritchie ME, Dunning MJ, Smith ML, Shi W, Lynch AG. 2011. BeadArray expression analysis using bioconductor. *PLoS Comput Biol* **7**(12): e1002276.

Rosa A, Brivanlou AH. 2011. A regulatory circuitry comprised of miR-302 and the transcription factors OCT4 and NR2F2 regulates human embryonic stem cell differentiation. *EMBO J* **30**(2): 237-248.

Rugg-Gunn PJ, Cox BJ, Ralston A, Rossant J. 2010. Distinct histone modifications in stem cell lines and tissue lineages from the early mouse embryo. *Proc Natl Acad Sci U S A* **107**(24): 10783-10790.

Sachs M, Onodera C, Blaschke K, Ebata KT, Song JS, Ramalho-Santos M. 2013. Bivalent chromatin marks developmental regulatory genes in the mouse embryonic germline in vivo. *Cell Rep* **3**(6): 1777-1784.

Salomonis N, Schlieve CR, Pereira L, Wahlquist C, Colas A, Zambon AC, Vranizan K, Spindler MJ, Pico AR, Cline MS et al. 2010. Alternative splicing regulates mouse embryonic stem cell pluripotency and differentiation. *Proc Natl Acad Sci U S A* **107**(23): 10514-10519.

Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of gene promoters. *Nature* **489**(7414): 109-113.

Schliep A, Schonhuth A, Steinhoff C. 2003. Using hidden Markov models to analyze gene expression time course data. *Bioinformatics* **19 Suppl 1**: i255-263.

Schoenfelder S, Sexton T, Chakalova L, Cope NF, Horton A, Andrews S, Kurukuti S, Mitchell JA, Umlauf D, Dimitrova DS et al. 2010. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet* **42**(1): 53-61.

Schoenherr CJ, Anderson DJ. 1995. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science* **267**(5202): 1360-1363.

Schuettengruber B, Cavalli G. 2009. Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. *Development* **136**(21): 3531-3542.

Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. 2008. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* **451**(7178): 535-540.

Sekido R, Bar I, Narvaez V, Penny G, Lovell-Badge R. 2004. SOX9 is up-regulated by the transient expression of SRY specifically in Sertoli cell precursors. *Dev Biol* **274**(2): 271-279.

Sekido R, Lovell-Badge R. 2008. Sex determination involves synergistic action of SRY and SF1 on a specific Sox9 enhancer. *Nature* **453**(7197): 930-934.

Sharan R, Ovcharenko I, Ben-Hur A, Karp RM. 2003. CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics* **19 Suppl 1**: i283-291.

Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, Crawford GE, Furey TS. 2013. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res* **23**(5): 777-788.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV et al. 2012. A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**(7409): 116-120.

Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford DK. 2014. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature biotechnology* **32**(2): 171-178.

Simon P. 2003. Q-Gene: processing quantitative real-time RT-PCR data. *Bioinformatics* **19**(11): 1439-1440.

Sinclair AH, Berta P, Palmer MS, Hawkins JR, Griffiths BL, Smith MJ, Foster JW, Frischauf AM, Lovell-Badge R, Goodfellow PN. 1990. A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature* **346**(6281): 240-244.

Sinha S, Adler AS, Field Y, Chang HY, Segal E. 2008. Systematic functional characterization of cis-regulatory motifs in human core promoters. *Genome Res* **18**(3): 477-488.

Small CL, Shima JE, Uzumcu M, Skinner MK, Griswold MD. 2005. Profiling gene expression during the differentiation and development of the murine embryonic gonad. *Biol Reprod* **72**(2): 492-501.

Smith AD, Sumazin P, Zhang MQ. 2005. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci U S A* **102**(5): 1560-1565.

Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**: Article3.

Smyth GK, Michaud J, Scott HS. 2005. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**(9): 2067-2075.

Sokalski KM, Li SK, Welch I, Cadieux-Pitre HA, Gruca MR, DeKoter RP. 2011. Deletion of genes encoding PU.1 and Spi-B in B cells impairs differentiation and induces pre-B cell acute lymphoblastic leukemia. *Blood* **118**(10): 2801-2808.

Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D et al. 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res* **21**(10): 1757-1767.

Soufi A, Donahue G, Zaret KS. 2012. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell* **151**(5): 994-1004.

Stevens NM. 1905. *Studies in Spermatogenesis with Especial Reference to the "Accessory Chromosome"*. Carnegie Institute of Washington, Washington D.C.

Stormo GD, Zhao Y. 2010. Determining the specificity of protein-DNA interactions. *Nat Rev Genet* **11**(11): 751-760.

Struhl K. 1998. Histone acetylation and transcriptional regulatory mechanisms. *Genes Dev* **12**(5): 599-606.

Suzuki H Forrest AR van Nimwegen E Daub CO Balwierz PJ Irvine KM Lassmann T Ravasi T Hasegawa Y de Hoon MJ et al. 2009. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet* **41**(5): 553-562.

Takahashi K, Yamanaka S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**(4): 663-676.

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* **96**(6): 2907-2912.

Thorrez L, Laudadio I, Van Deun K, Quintens R, Hendrickx N, Granvik M, Lemaire K, Schraenen A, Van Lommel L, Lehnert S et al. 2011. Tissue-specific disallowance of housekeeping genes: the other face of cell differentiation. *Genome Res* **21**(1): 95-105.

Tijssen MR, Cvejic A, Joshi A, Hannah RL, Ferreira R, Forrai A, Bellissimo DC, Oram SH, Smethurst PA, Wilson NK et al. 2011. Genome-wide analysis of simultaneous

GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev Cell* **20**(5): 597-609.

Tse E, Smith AJ, Hunt S, Lavenir I, Forster A, Warren AJ, Grutz G, Foroni L, Carlton MB, Colledge WH et al. 2004. Null mutation of the Lmo4 gene or a combined null mutation of the Lmo1/Lmo3 genes causes perinatal lethality, and Lmo4 controls neural tube development in mice. *Mol Cell Biol* **24**(5): 2063-2073.

Vainio S, Heikkila M, Kispert A, Chin N, McMahon AP. 1999. Female development in mammals is regulated by Wnt-4 signalling. *Nature* **397**(6718): 405-409.

Vavouri T, Elgar G. 2005. Prediction of cis-regulatory elements using binding site matrices--the successes, the failures and the reasons for both. *Curr Opin Genet Dev* **15**(4): 395-402.

Venkatesh S, Smolle M, Li H, Gogol MM, Saint M, Kumar S, Natarajan K, Workman JL. 2012. Set2 methylation of histone H3 lysine 36 suppresses histone exchange on transcribed genes. *Nature* **489**(7416): 452-455.

Vidal VP, Chaboissier MC, de Rooij DG, Schedl A. 2001. Sox9 induces testis development in XX transgenic mice. *Nat Genet* **28**(3): 216-217.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**(7231): 854-858.

Visel A, Taher L, Girgis H, May D, Golonzhka O, Hoch RV, McKinsey GL, Pattabiraman K, Silberberg SN, Blow MJ et al. 2013. A high-resolution enhancer atlas of the developing telencephalon. *Cell* **152**(4): 895-908.

Walker MD, Edlund T, Boulet AM, Rutter WJ. 1983. Cell-specific expression controlled by the 5'-flanking region of insulin and chymotrypsin genes. *Nature* **306**(5943): 557-561.

Wasserman WW, Fickett JW. 1998. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* **278**(1): 167-181.

Western PS, Ralli RA, Wakeling SI, Lo C, van den Bergen JA, Miles DC, Sinclair AH. 2011. Mitotic arrest in teratoma susceptible fetal male germ cells. *PLoS One* **6**(6): e20736.

Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**(2): 307-319.

Wilhelm D, Hiramatsu R, Mizusaki H, Widjaja L, Combes AN, Kanai Y, Koopman P. 2007. SOX9 regulates prostaglandin D synthase gene transcription in vivo to ensure testis development. *J Biol Chem* **282**(14): 10553-10560.

Wilson NK, Foster SD, Wang X, Knezevic K, Schutte J, Kaimakis P, Chilarska PM, Kinston S, Ouwehand WH, Dzierzak E et al. 2010. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7**(4): 532-544.

Wu C, Gilbert W. 1981. Tissue-specific exposure of chromatin structure at the 5' terminus of the rat preproinsulin II gene. *Proc Natl Acad Sci U S A* **78**(3): 1577-1580.

Xu J, Shao Z, Glass K, Bauer DE, Pinello L, Van Handel B, Hou S, Stamatoyannopoulos JA, Mikkola HK, Yuan GC et al. 2012. Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev Cell* **23**(4): 796-811.

Yamashita R, Suzuki Y, Sugano S, Nakai K. 2005. Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene* **350**(2): 129-136.

Yao HH, Matzuk MM, Jorgez CJ, Menke DB, Page DC, Swain A, Capel B. 2004. Follistatin operates downstream of Wnt4 in mammalian ovary organogenesis. *Dev Dyn* **230**(2): 210-215.

Yordy JS, Moussa O, Pei H, Chaussabel D, Li R, Watson DK. 2005. SP100 inhibits ETS1 activity in primary endothelial cells. *Oncogene* **24**(5): 916-931.

Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J, Jonsdottir GA, Ruotti V, Stewart R et al. 2007. Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**(5858): 1917-1920.

Yuan M, Kendziorski C. 2006. Hidden Markov models for microarray time course data in multiple biological conditions. *Journal of the American Statistical Association* **101**(476): 1323-1332.

Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. 2009. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**(15): 1952-1958.

Zaret KS, Carroll JS. 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* **25**(21): 2227-2241.

Zeng PY, Vakoc CR, Chen ZC, Blobel GA, Berger SL. 2006. In vivo dual cross-linking for identification of indirect DNA-associated proteins by chromatin immunoprecipitation. *Biotechniques* **41**(6): 694, 696, 698.

Zhang Z, Verheyden JM, Hassell JA, Sun X. 2009. FGF-regulated Etv genes are essential for repressing Shh expression in mouse limb buds. *Dev Cell* **16**(4): 607-613.

Zhu J, He F, Hu S, Yu J. 2008. On the nature of human housekeeping genes. *Trends Genet* **24**(10): 481-484.

Zwart W, Koornstra R, Wesseling J, Rutgers E, Linn S, Carroll JS. 2013. A carrier-assisted ChIP-seq method for estrogen receptor-chromatin interactions from breast cancer core needle biopsy samples. *BMC Genomics* **14**: 232.

Biography

Anirudh Natarajan was born on March 17th, 1984 in Jamshedpur, India to Vijayalakshmi and Subramaniam Natarajan. Until he was 13, he lived in Madras (Chennai), India with his parents and brother. Apart from going to school, he remembers watching hour-long standoffs between their dogs and the neighborhood cobras. In the 8th grade, he moved with his parents to Jakarta, Indonesia where he attended Gandhi Memorial International School. He graduated in May 2001, and began his Bachelor's degree in Electrical Engineering in the Indian Institute of Technology, Madras. After a year of troubling personal and academic circumstances, he started his Bachelor's in Electrical Engineering in the National University of Singapore. He had a far better experience in Singapore, with the icing on the cake being that he met his now wife, Lydia Ng, while there. He graduated with a Bachelor's degree in 2006 and went on to conduct research in cellular and Body Area Networks, with his undergraduate thesis advisors, Dr. Mehul Motani and Dr. Vikram Srinivasan. After a couple years, he felt he wanted to be a scientist rather than an engineer, and when someone mentioned computational biology to him, he borrowed his now wife's Molecular Biology of the Cell by Alberts et al. Despite not having touched biology in a decade, by the time he read the chapter on DNA replication, he was hooked. With Lydia's support and encouragement, he applied to graduate school and was accepted into the Program in Computational

Biology and Bioinformatics at Duke University in 2009. He joined the Capel and Ohler labs in the summer of 2010. He and Lydia, his partner/motivator/enabler of 6 and half years, got hitched on July 21st, 2012.

Publications (* indicates co-first authorship)

Munger, S.C.*, Natarajan, A.*, Looger, L.L., Ohler, U., and Capel, B. (2013) Fine Time Course Expression Analysis Identifies Cascades of Activation and Repression and Maps a Regulator of Mammalian Sex Determination. PLoS Genetics 9(7): e1003630

Natarajan, A., Yardimci, G., Sheffield, N., Crawford, G.E., and Ohler, U. (2012) Predicting Cell-Type Specific Gene Expression from Regions of Open Chromatin, Genome Research September 2012 22: 1711-1722.

Jameson, S.A., Natarajan, A., Cool, J., DeFalco, T., Maatouk, D.M., Mork, L., Munger, S.C., Capel, B. (2011) Temporal Transcriptional Profiling of Somatic and Germ Cells Reveals Biased Lineage Priming of Sexual Fate in the Fetal Mouse Gonad, PLoS Genetics 8(3): e1002575

Previous publications in Engineering

Natarajan, A., de Silva, B., Yap, K-K., Motani, M. (2009) Link layer behavior of body area networks at 2.4 GHz. In Proceedings of MobiCom 2009.

Natarajan, A., de Silva, B., Yap, K-K., Motani, M. To Hop or Not to Hop: Network Architecture for Body Sensor Networks. In Proceedings of SECON 2009.

de Silva, B., Natarajan, A., Motani, M., Chua, K-C., Inter-User Interference in Body Sensor Networks: Preliminary Investigation and an Infrastructure Based Solution. In Proceedings of BSN 2009.

de Silva B., Natarajan A., Motani, M., Chua, K-C., A Real-Time Exercise Feedback Utility with Body Sensor Networks. In Proceedings of BSN 2008.

de Silva B., Natarajan A., Motani, M., Chua, K-C., Design Consideration of Body Sensor Networks. In Proceedings of Healthcom 2008.

Motani M., Yap K-K., Natarajan, A., de Silva, B., Siquan, H., Chua, K-C., Network Characteristics of Urban Environments in Wireless BAN. In Proceedings of Healthcom 2008.

Aziz, N.T., Khowaja, F., Natarajan, A., Motani, M., Srinivasan, V., FluLog: An Automated Contact Tracing Tool for Mitigating Pandemic Spread. In Proceedings of WSNHC 2007.

Natarajan, A., Motani, M., de Silva, B., Yap, K-K., Chua, K-C., Investigating Network Architectures for Body Sensor Networks. In Proceedings of HealthNet 2007.

Natarajan, A., Motani, M., Srinivasan, V., Understanding Urban Interactions from Bluetooth Phone Contact Traces. In Proceedings of PAM 2007.